

Steady-state Diffusion Approximations for Discrete-time Queue in Hospital Inpatient Flow Management

Jiekun Feng

Department of Statistical Science, Cornell University

Pengyi Shi

Krannert School of Management, Purdue University

Abstract

In this paper, we analyze a discrete-time queue that is motivated from studying hospital inpatient flow management, where the customer count process captures the midnight inpatient census. The stationary distribution of the customer count has no explicit form and is difficult to compute in certain parameter regimes. Using the Stein's method framework, we identify a continuous random variable to approximate the steady-state customer count. The continuous random variable corresponds to the stationary distribution of a diffusion process with *state-dependent* diffusion coefficients. We characterize the error bounds of this approximation under a variety of system load conditions – from lightly loaded to heavily loaded. We also identify the critical role that the service rate plays in the convergence rate of the error bounds. We perform extensive numerical experiments to support the theoretical findings and to demonstrate the approximation quality. In particular, we show that our approximation performs better than those based on constant diffusion coefficients when the number of servers is small.

Keywords: Discrete Queue, Steady-state Analysis, Stein's Method, State-dependent Diffusion

1. Introduction

In this paper, we analyze a $GI/Geo/N$ discrete-time queue, or *discrete queue* in short. This discrete queue has N identical servers and a buffer that can hold infinitely many customers. Customer arrivals and departures occur at discrete time epochs $k = 0, 1, 2, \dots$. At each epoch k , a total

Email addresses: jf646@cornell.edu (Jiekun Feng), shi178@purdue.edu (Pengyi Shi)

5 number of D_k customers depart from the queue first, and then a total number of A_k customers arrive. If there are enough servers, we admit all waiting customers (if any) and new arrivals into service; otherwise, we admit as many customers as possible, following the first-come-first-served queueing discipline, until all servers are occupied and hold the remaining customers in the buffer. For the arrival process, we assume that $\{A_k, k = 0, 1, \dots\}$ forms a sequence of independent and
10 identically distributed (i.i.d.) random variables. For the departure process, we assume that each customer in service at the beginning of epoch k (excluding new arrivals) has a constant probability $\mu \in (0, 1)$ of departing in epoch k . It is equivalent to assuming that the “service time” in this discrete queue follows a *geometric* distribution with the success parameter being μ , which is why we use “Geo” in the notation of the queue; see [1] for a more rigorous proof on this equivalence
15 using a coupling argument.

This discrete queue is motivated from studying the inpatient midnight census in hospitals [1], where the servers correspond to inpatient beds, and customer arrivals and departures correspond to patient bed-requests and discharges in a day; see more motivation in Section 1.3. We focus on the customer count process and analyze its steady-state performance. Let X_k denote the total number of customers in system at the beginning of epoch k , including both the customers in service and those waiting in the buffer. Under our arrival and departure assumptions, the customer count process $X = \{X_k : k = 0, 1, \dots\}$ forms a discrete-time Markov chain (DTMC) and is characterized by the following relationship:

$$X_{k+1} = X_k + A_k - D_k, \quad k = 0, 1, \dots \quad (1.1)$$

Here, the total number of departures D_k follows a binomial distribution with parameters (Z_k, μ) , where $Z_k \equiv X_k \wedge N$ is the number of busy servers at the beginning of epoch k with \wedge denoting the minimum between two real numbers. In the rest of this paper, we focus on the *Poisson* arrival case, that is, A_k follows a common Poisson distribution with mean Λ . We specify the treatments
20 for general arrival distributions in an online supplement [2] and establish the corresponding error bounds there to keep this paper focused.

When

$$\Lambda < N\mu, \quad \text{or equivalently,} \quad R \equiv \frac{\Lambda}{\mu} < N, \quad (1.2)$$

the DTMC X has a unique stationary distribution π [1]. Here, R is the *offered load* of the discrete queue. We use X_∞ to denote the steady-state customer count, and correspondingly, its distribution

is π . We also define a *scaled* version of X_∞ as

$$\tilde{X}_\infty = (X_\infty - R)/\sqrt{R}. \quad (1.3)$$

1.1. Results summary

We identify a continuous random variable (r.v.) Y_∞ to approximate the scaled steady-state customer count \tilde{X}_∞ . We establish bounds on the approximation errors under various system conditions. Specifically, Y_∞ is defined by the density

$$p(x) = \frac{\kappa}{a(x)} \exp \left(\int_0^x \frac{2b(y)}{a(y)} dy \right), \quad x \in \mathbb{R}, \quad (1.4)$$

where $\kappa > 0$ is a normalizing constant such that $\int_{-\infty}^{\infty} p(x) dx = 1$,

$$b(x) \equiv \mu [(x + \zeta)^- - \zeta^-] = \begin{cases} -\mu x, & x \leq -\zeta, \\ \mu \zeta, & x \geq -\zeta, \end{cases} \quad (1.5)$$

$$a(x) \equiv \begin{cases} \mu(1 + \Lambda), & x \leq -\sqrt{R}, \\ \mu(2 - \mu + \delta(1 - \mu)x + \mu x^2), & x \in [-\sqrt{R}, -\zeta], \\ \mu(2 - \mu + \delta(1 - \mu)|\zeta| + \mu \zeta^2), & x \geq -\zeta, \end{cases} \quad (1.6)$$

and

$$\zeta \equiv (R - N)/\sqrt{R} < 0. \quad (1.7)$$

For any given $N \geq 1$, $\Lambda > 0$, and $\mu \in (0, 1)$ such that $1 \leq R < N$ and

$$\mu = \gamma R^{-s}, \quad N - R = \beta R^q \quad (1.8)$$

for some constants $\gamma, \beta > 0$ and $s, q \geq 0$, we establish the following error bounds on the Wasserstein distance between \tilde{X}_∞ and Y_∞ , defined as

$$d_W(\tilde{X}_\infty, Y_\infty) \equiv \sup_{h \in \text{Lip}(1)} |\mathbb{E}h(\tilde{X}_\infty) - \mathbb{E}h(Y_\infty)| \quad (1.9)$$

with $\text{Lip}(1) = \{h : \mathbb{R} \rightarrow \mathbb{R}, |h(x) - h(y)| \leq |x - y| \text{ for all } x, y \in \mathbb{R}\}$. Note that convergence in the Wasserstein distance implies convergence in distribution [3].

Theorem 1. For any $s \in [1/2, 1)$ and $q \in [1/2, 1]$ such that (1.8) is satisfied,

$$d_W(\tilde{X}_\infty, Y_\infty) \leq C_1(\gamma, \beta) R^{-s/2}. \quad (1.10)$$

Theorem 2. For any $s \geq 1$ and $q \in [0, 1]$ such that (1.8) is satisfied and $R \geq 2\gamma$,

$$d_W(\tilde{X}_\infty, Y_\infty) \leq C_2(\gamma, \beta) R^{-1/2}. \quad (1.11)$$

Here, $C_1(\gamma, \beta)$ and $C_2(\gamma, \beta)$ are two constants that only depend on γ and β and can be recovered in the proof. Note that s characterizes the rate of μ converging to 0, and q characterizes the system load condition, with $q = 0, 1/2, 1$ corresponding to the non-degenerate slowdown (NDS) [4], quality-and-efficiency-driven (QED), and quality-driven (QD) regimes [5], respectively. We give more interpretations of these two theorems below in Section 1.2.

1.2. Main contributions

Comparing to the recent series of papers on Stein's method for steady-state approximation, our paper makes the following contributions.

- The density of Y_∞ corresponds to the steady-state distribution of a diffusion process with a *state-dependent* diffusion coefficient $a(x)$ and a piece-wise linear drift $b(x)$. Note that in [1], the most relevant paper, the authors studied the same DTMC X and promoted the use of a continuous r.v. with similar state-dependent diffusion coefficients to approximate \tilde{X}_∞ . However, they were not able to establish the error bounds for such approximation – Theorem 3 there only established the error bounds between \tilde{X}_∞ and a continuous r.v. with a *constant* diffusion coefficient, which we denote as Y_∞^0 in this paper. Comparing to Y_∞^0 , Y_∞ is equally easy to evaluate numerically using the explicit form in (1.4), while it produces better approximation for \tilde{X}_∞ , especially when the system size is small. For example, when $N = 18$, Y_∞ can reduce the relative approximation error in the expected queue length by as much as 10% than using Y_∞^0 . This paper fills the gap in establishing the error bounds between \tilde{X}_∞ and the state-dependent r.v. Y_∞ . As we will further illustrate below, this is not a trivial extension – it is challenging to establish bounds involving the density of Y_∞ since its form is more complicated than that of Y_∞^0 .
- We characterize the convergence rate of the error bounds under different system load conditions (reflected by q) and rates of μ converging to 0 (reflected by s). Theorem 1 says that

when μ converges to 0 in a rate that is between $1/\sqrt{R}$ and $1/R$, and when the system runs in the QD, QED, or any regime in between, the error bound converges to 0 in a rate that is between $1/\sqrt[4]{R}$ and $1/\sqrt{R}$. Theorem 2 says that when μ converges to 0 in a rate that is faster than or equal to $1/R$, and when the system runs in any regimes from QD to NDS, the error bound converges to 0 at a *constant* rate $1/\sqrt{R}$. Comparing to Theorem 3 in [1] that only established the error bounds in the QED regime, our results here cover a much wider range of q , which justify using Y_∞ to approximate X_∞ under a variety of system load conditions. This can have more practical impact since, for example, the utilization of different hospital wards can vary greatly from lightly loaded (60%) to very heavily loaded ($> 95\%$); see for example, Table 5 in [6], and Tables 89 and 91 in [7].

- Theorems 1 and 2 also reveal the critical role that the service rate μ plays in the convergence of the error bound in the discrete queue. That is, to ensure that the error bound goes to 0, we require that the number of servers N goes to ∞ and the service rate μ goes to 0 *at the same time*. This is a main difference from the continuous-time queueing systems studied in the literature, including [8], where the authors develop a state-dependent diffusion model for Erlang-C queue. In the continuous-time queues, just $N, R \rightarrow \infty$ ensures the convergence of the error bound. In addition, in the discrete queue the rate of μ converging to 0 needs to be *fast* enough to ensure convergence in different operating regimes, e.g., in the NDS regime, our numerical results show that the error bound does *not* converge when $s = 1/2$, but does converge when $s = 1$; see Section 4. Note that this asymptotic regime of μ going to 0 is also of practical relevance for hospital setting since a typical inpatient stay (service time) is 4-5 days [1], i.e., the service rate is small.
- For the proof, we overcome two major challenges in establishing the error bounds. The first one is that, due to the state-dependent diffusion coefficient, the density $p(x)$ becomes very complicated. Thus, the gradient bounds, which involves integrals with $p(x)$, cannot be directly estimated as in the constant diffusion case. The second one is that, unlike our continuous-time counterpart – the Erlang-C queue – where the customer count process is a birth-death process with only one arrival or departure occurring at each transition, in the discrete queue a number of arrivals and discharges could occur at each transition. The latter makes the proof of the moment bounds complicated, especially when the system is heavily loaded (NDS

regime). See more details in Section 3.

80 A last thing worth noting here is that there is no explicit formula to calculate π despite the simplicity of the dynamic equation (1.1). Using the standard Markov chain technique to solve π can be time consuming since generating the transition matrix requires calculation of the convolution between A_k and D_k , where the distribution of the latter depends on the number of customers X_k . Even for realistic hospital settings, it could take several hours, sometimes days, to get π ;
85 see computational results summarized in Table B.7. In addition, because the distribution of D_k depends on X_k instead of following a common distribution (which differentiates Equation (1.1) from the dynamics of the $M/GI/1$ queue), the conventional generating-function method does not work well either; see more in Section 1.4. In contrast, the density function of Y_∞ has an explicit form, which provides an efficient engineering tool to approximately calculate π .

90 The rest of the paper is organized as follows. In Section 1.3, we discuss the motivation of the discrete queue from hospital inpatient flow management and its broader application in telecommunication. In Section 1.4 we review relevant papers in the literature. In Section 2 we prove our main theorems. In Section 3, we detail the proof for two important lemmas to establish the error bounds. We describe the numerical results in Section 4 and conclude the paper in Section 5.

95 1.3. Motivation of the discrete queue

The discrete queue is motivated by studying hospital inpatient flows [10]. The inpatient beds are modeled as the servers, and patients who need to be admitted to an inpatient bed are modeled as customers, for example, patients who have received treatment in the emergency department (ED) and wait to be hospitalized – commonly known as the ED boarding patients. The customer count
100 X_k corresponds to the *midnight* census at day k , i.e., the number of patients who are occupying an inpatient bed or waiting to be admitted at the midnight of day k . Naturally, A_k and D_k correspond to the total number of patient arrivals and discharges within day k , respectively, and the midnight census at the next day, X_{k+1} , evolves as in (1.1). Empirical studies suggest that the bed-request process of the ED boarding patients can be modeled by a periodic Poisson process with the period
105 being one day [10, 11]. Thus, it is reasonable to assume that the daily arrival A_k follows a Poisson distribution, with Λ corresponding to the daily arrival rate. However, the length-of-stay (LOS) distribution is usually *not* geometric. Nevertheless, the system performance is not very sensitive to

the LOS distribution when the utilization is not extremely high; see Section 4.7 of [12]. Therefore, we focus on the geometric setting in this paper for tractability.

110 The midnight census is a key performance metric monitored by many hospitals [13]. Moreover, getting the stationary distribution for the midnight census is a crucial step in predicting the time-of-day patient census, since the census at a certain hour t equals the sum of the midnight census and the difference between the number of arrivals and discharges from the midnight to hour t ; see the two-time-scale framework developed in [1]. Also see there for more details on the importance
115 of studying the midnight census as well as further justifications on model assumptions.

Besides the applications in the healthcare setting, discrete queueing systems have been motivated from a variety of applications in the fields of telecommunication and computer systems, in which the time is usually divided into fixed-length time slots. For example, multi-server discrete queues with geometric service times are studied in the context of circuit-switched multiple-access
120 communications channel [14] and systems with randomly interrupted servers [15] (which are equivalent to systems with non-interrupted servers and geometric service time). Thus, the analysis and results we gain in this paper can potentially benefit a larger community.

1.4. Literature review

Stein's method is a well known method for establishing error bounds in various fields and
125 applications; see, for example, the survey papers [16, 17]. The proof in our paper is mainly based on the framework developed in [18] on applying Stein's method to steady-state diffusion approximations in $M/Ph/N + M$ queue; also see a tutorial on applying this framework to Erlang-A and Erlang-C queues in [19] and the references there for this line of work. In [8], the authors develop a new diffusion model with state-dependent diffusion coefficients for Erlang-C queue; this refined diffusion
130 model allows them to establish an error bound with a higher order convergence rate. Indeed, their paper is a main driver for us to look into the state-dependent diffusion approximation in the discrete queue setting, and the proof there has helped us in establishing the gradient bounds in Section 3.2. In addition to error bounds on the customer count distributions, Gurvich and Huang [20] apply the Stein's method framework to the $M/GI/1 + GI$ queue and establish error bounds for a variety
135 of performance measures including the waiting time and abandonment. We have specified the differences of our paper and these papers in the contribution part above.

Regarding the modeling side, various discrete queues have been studied in the area of telecom-

munication and computer systems, with different assumptions on the arrival and service time distribution, server numbers (single or multiple), and buffer capacity; see a detailed summary in
140 Section 1.2.7 of [21] for some of the early works and [22] for more recent development. The most relevant work is [22], where the authors study a same $GI/Geo/N$ queue. In that paper, the authors employ the conventional generating-function method to perform steady-state analysis. However, numerically implementing the generating-function method to find π involves finding $N - 1$ roots inside the unit disk from an N th order nonlinear equation, and it becomes computationally difficult
145 when N is large. Indeed, in their numerical experiments, the largest N tried by the authors is 16. In this paper, we develop an efficient and accurate way to approximate π , and more importantly, we are able to provide error bounds on such approximation. Janssen et al. [23] study the $M/D/s$ queue where the dynamic equation for the number of waiting customers has a similar form to (1.1), but intrinsically is different. Their objective is similar to ours, that is, finding more accurate approx-
150 imations for performance metrics of interest, with a focus on the QED regime. The method used in [23] involves infinite series expansion, which requires detailed derivations tied to the queueing models, performance metrics, and operating regimes; see [19] for more discussions on comparing this method with the Stein's method framework.

2. Proof of Theorems 1 and 2

To prove Theorems 1 and 2, we employ the Stein's method framework for steady-state approximation. The major components of this framework are Poisson equation, generator coupling, gradient and moment bounds; see [18] for a systematical description. Throughout the rest of the paper, we define

$$\delta \equiv 1/\sqrt{R}$$

155 for notational convenience.

Proof. Define G_Y as

$$G_Y f(x) = b(x)f'(x) + \frac{1}{2}a(x)f''(x), \quad x \in \mathbb{R}, f \in C^2(\mathbb{R}), \quad (2.1)$$

which is the generator of a diffusion process with diffusion coefficient $a(x)$ and drift $b(x)$. As noted before, the stationary distribution of this diffusion process has a density given by (1.4).

Now, let $f = f_h$ be a solution to the *Poisson equation*

$$G_Y f(x) = \mathbb{E}[h(Y_\infty)] - h(x), \quad x \in \mathbb{R}. \quad (2.2)$$

Lemma 2 below shows that f is twice continuously differentiable, with an absolutely continuous second derivative.

Next, we do a *generator coupling* via (2.2). The generator of the scaled DTMC \tilde{X} is

$$G_{\tilde{X}} f(x) = \mathbb{E}_n[f(x + \delta(A_0 - D_0)) - f(x)], \quad x = \delta(n - x_\infty), \quad n = 0, 1, \dots, \quad f \in C^2(\mathbb{R}), \quad (2.3)$$

where \mathbb{E}_n is the expectation under \mathbb{P}_n , the conditional probability distribution given that the starting customer count equals n , $A_0 \sim \text{Poisson}$ with mean Λ , and $D_0 \sim \text{binomial}$ with $(n \wedge N, \mu)$ and is independent of A_0 . It is proven in [1] that

$$\mathbb{E}[G_{\tilde{X}} f(\tilde{X}_\infty)] = 0. \quad (2.4)$$

160 Taking expectation with respect to \tilde{X}_∞ on both sides of (2.2), we have that

$$\begin{aligned} \mathbb{E}[h(Y_\infty)] - \mathbb{E}[h(\tilde{X}_\infty)] &= \mathbb{E}[G_Y f(\tilde{X}_\infty)] \\ &= \mathbb{E}[G_Y f(\tilde{X}_\infty) - G_{\tilde{X}} f(\tilde{X}_\infty)], \end{aligned} \quad (2.5)$$

where the second equality comes from (2.4). To bound the right side of (2.5), we perform the Taylor expansion for $G_{\tilde{X}} f(x)$, for any given $x = \delta(n - x_\infty)$ and $n = 0, 1, \dots$,

$$\begin{aligned} G_{\tilde{X}} f(x) &= f'(x)\delta\mathbb{E}_n(A_0 - D_0) + \frac{1}{2}f''(x)\delta^2\mathbb{E}_n[(A_0 - D_0)^2] + \frac{1}{2}\delta^2\mathbb{E}_n[(f''(\eta) - f''(x))(A_0 - D_0)^2] \\ &= G_Y f(x) + \frac{1}{2}\delta^2\mathbb{E}_n\left[\int_x^\eta f'''(y)dy(A_0 - D_0)^2\right], \end{aligned} \quad (2.6)$$

with

$$|\eta - x| \leq \delta|A_0 - D_0|.$$

Note that (2.6) follows from the absolute continuity of $f''(x)$ and the following facts:

$$\begin{aligned} \delta\mathbb{E}_n(A_0 - D_0) &= \delta[\Lambda - (n \wedge N)\mu] = \delta(\Lambda - N\mu) + \delta(n - N)^-\mu \\ &= \mu\zeta + (x + \zeta)^-\mu = b(x), \end{aligned} \quad (2.7)$$

$$\begin{aligned} \delta^2\mathbb{E}_n[(A_0 - D_0)^2] &= \delta^2\text{Var}_n(A_0 - D_0) + \delta^2(\mathbb{E}_n(A_0 - D_0))^2 \\ &= \delta^2\Lambda + \delta^2(n \wedge N)\mu(1 - \mu) + b^2(x) = a(x). \end{aligned} \quad (2.8)$$

Note that in (1.6), the part of $a(x) = \mu(1 + \Lambda)$ on $(-\infty, -1/\delta)$ corresponds to $n < 0$. It is added to make the diffusion coefficient $a(x)$ exist everywhere on \mathbb{R} and be continuous at the point $x = -1/\delta$, so that the corresponding diffusion process is well-defined. Here, to derive the second equality of (2.8), we have used the fact that A_0 is Poisson so that its variance equals the mean Λ . The treatment for the non-Poisson arrival distribution and the corresponding error bounds are detailed in the online supplement [2].

Combining (2.5) and (2.6) gives

$$\begin{aligned} \left| \mathbb{E}h(\tilde{X}_\infty) - \mathbb{E}h(Y_\infty) \right| &= \left| \mathbb{E}[G_{\tilde{X}}f(x) - G_Yf(x)] \right| \\ &\leq \frac{1}{2}\delta^2\mathbb{E}[\epsilon(X_\infty)], \end{aligned} \quad (2.9)$$

where

$$\epsilon(X_\infty) = \mathbb{E}_{X_\infty} \left[\left| \int_{\tilde{X}_\infty}^{\tilde{X}_\infty + \delta(A_0 - D_0)} |f'''(y)| dy \right| (A_0 - D_0)^2 \right], \quad (2.10)$$

and we use \mathbb{E}_{X_∞} to denote the expectation conditioning on the starting customer count X_∞ in the rest of the paper. Then, Theorems 1 and 2 follow from part (a) and part (b) of Lemma 1 below, respectively. \square

Lemma 1. (a) For an $s \in [1/2, 1]$ and any $q \in [1/2, 1]$ such that (1.8) is satisfied,

$$\frac{1}{2}\delta^2\mathbb{E}[\epsilon(X_\infty)] \leq C_3(\gamma, \beta)R^{-s/2}. \quad (2.11)$$

(b) For any $s \geq 1$ and any $q \in [0, 1]$ such that (1.8) is satisfied and $R \geq 2\gamma$,

$$\frac{1}{2}\delta^2\mathbb{E}[\epsilon(X_\infty)] \leq C_4(\gamma, \beta)R^{-1/2}. \quad (2.12)$$

Here, $C_3(\gamma, \beta)$ and $C_4(\gamma, \beta)$ are two constants depending only on γ, β , and can be easily recovered from the proof detailed in Section 3. Note that the condition $R \geq 2\gamma$ in part (b) implies $\mu \leq 1/2$, which is a realistic assumption for hospital inpatient setting since the typical service time is longer than 2 days. However, this condition can be relaxed – we can prove the same result as long as $0 < \mu \leq \mu_0 < 1$ for some constant μ_0 . We focus on the special case of $R \geq 2\gamma$ to keep the statement of the results clean.

To prove Lemma 1, we also need the following lemma on the gradient bounds, whose proof is also detailed in Section 3.

Lemma 2. Fix an $h \in \text{Lip}(1)$ with $h(0) = 0$. There exists a solution to the Poisson equation (2.2), f_h , that is twice continuously differentiable with an absolutely continuous second derivative, for any $s \geq 1/2$ such that (1.8) is satisfied,

$$|f_h'''(x)| \leq \begin{cases} \frac{C_0}{\mu}(1 + 1/|\zeta|), & x \leq -\zeta, \\ \frac{4}{\mu}, & x \geq -\zeta, \end{cases} \quad (2.13)$$

180 where $f_h'''(x)$ is interpreted as the left derivative at the point $x = -1/\delta$ and $x = -\zeta$, and $C_0 = C_0(\gamma)$ is a constant depending only on γ .

3. Proof outline of Lemmas 1 and 2

As mentioned in the introduction, we overcome two major challenges in establishing the error bounds. The first one is to establish the gradient bound in Lemma 2 when the density $p(x)$ has a complicated form; and the second one is to bound the error term in Lemma 1. In the interest of space, we outline some of the key points in proving these two lemmas in Sections 3.2 and 3.1, while leaving the complete details to Section 1 of the online supplement [2].

3.1. Lemma 1

The difficulty in proving Lemma 1 comes from the fact that the discrete queue could have multiple arrivals and departures between two contiguous epochs, in other words, the jump size between X_k and X_{k+1} is unbounded. The proof involves two key steps: one is to bound $\mathbb{E}_{X_\infty} [|A_0 - D_0|^3]$ and the other is to establish a bound on the idle probability $\mathbb{P}(X_\infty \leq N)$ with a *coupling* argument. To see this, first consider when $X_\infty \leq N$, or equivalently, $\tilde{X}_\infty \leq -\zeta$,

$$\begin{aligned} \epsilon(X_\infty) &= \mathbb{E}_{X_\infty} \left[\left| \int_{\tilde{X}_\infty}^{\tilde{X}_\infty + \delta(A_0 - D_0)} |f'''(y)| dy \right| (A_0 - D_0)^2 \mathbf{1}_{\{X_\infty + A_0 - D_0 \leq N\}} \right] \\ &\quad + \mathbb{E}_{X_\infty} \left[\left| \int_{\tilde{X}_\infty}^{\tilde{X}_\infty + \delta(A_0 - D_0)} |f'''(y)| dy \right| (A_0 - D_0)^2 \mathbf{1}_{\{X_\infty + A_0 - D_0 > N\}} \right] \\ &\leq \mathbb{E}_{X_\infty} \left[\left| \int_{\tilde{X}_\infty}^{\tilde{X}_\infty + \delta(A_0 - D_0)} \frac{C_0}{\mu} (1 + 1/|\zeta|) dy \right| (A_0 - D_0)^2 \mathbf{1}_{\{X_\infty + A_0 - D_0 \leq N\}} \right] \\ &\quad + \mathbb{E}_{X_\infty} \left[\left| \int_{\tilde{X}_\infty}^{-\zeta} \frac{C_0}{\mu} (1 + 1/|\zeta|) dy \right| (A_0 - D_0)^2 \mathbf{1}_{\{X_\infty + A_0 - D_0 > N\}} \right] \\ &\quad + \mathbb{E}_{X_\infty} \left[\left| \int_{-\zeta}^{\tilde{X}_\infty + \delta(A_0 - D_0)} \frac{4}{\mu} dy \right| (A_0 - D_0)^2 \mathbf{1}_{\{X_\infty + A_0 - D_0 > N\}} \right] \end{aligned}$$

$$= \delta \frac{C_0}{\mu} (1 + 1/|\zeta|) \mathbb{E}_{X_\infty} \left[|(X_\infty + A_0 - D_0) \wedge N - X_\infty| (A_0 - D_0)^2 \right] \quad (3.1)$$

$$+ \delta \frac{4}{\mu} \mathbb{E}_{X_\infty} \left[(A_0 - D_0 - (N - X_\infty))(A_0 - D_0)^2 \mathbf{1}_{\{X_\infty + A_0 - D_0 > N\}} \right]. \quad (3.2)$$

The first inequality comes from the gradient bound in Lemma 2, and we separately consider the two cases $X_\infty + A_0 - D_0 > N$ and $X_\infty + A_0 - D_0 \leq N$ since the gradient bounds for f''' are different on $(-\infty, -\zeta]$ and $[-\zeta, \infty)$. After some algebra, we get

$$\begin{aligned} & \frac{1}{2} \delta^2 \mathbb{E} [\epsilon(X_\infty) \mathbf{1}_{\{X_\infty \leq N\}}] \\ & \leq \delta^3 \frac{2}{\mu} \mathbb{E} [\mathbb{E}_{X_\infty} |A_0 - D_0|^3] + \frac{C_0}{2} \delta^3 \frac{1}{\mu} (1 + 1/|\zeta|) \mathbb{E} [\mathbb{E}_{X_\infty} |A_0 - D_0|^3 \mathbf{1}_{\{X_\infty \leq N\}}]. \end{aligned} \quad (3.3)$$

Similarly, for $X_\infty > N$, we get

$$\begin{aligned} & \frac{1}{2} \delta^2 \mathbb{E} [\epsilon(X_\infty) \mathbf{1}_{\{X_\infty > N\}}] \\ & \leq \delta^3 \frac{2}{\mu} \mathbb{E} [\mathbb{E}_{X_\infty} |A_0 - D_0|^3] + \frac{C_0}{2} \delta^3 \frac{1}{\mu} (1 + 1/|\zeta|) \mathbb{E} [|A - D|^3] \mathbb{P}(X_\infty \leq N), \end{aligned} \quad (3.4)$$

where $A \sim \text{Poisson}(\Lambda)$ and $D \sim \text{binomial}(N, \mu)$ are independent of X_∞ .

190 For part (a) of Lemma 1, we need a special treatment to bound $\mathbb{E}_{X_\infty} [|A_0 - D_0|^3]$ since it is in the order of μN , and under the condition $s \in [1/2, 1)$, μN increases to ∞ when $N \rightarrow \infty$. This treatment is similar to the technique used for proving Lemma 6 in [1], and thus, the details are omitted here. Once we have bounded $\mathbb{E}_{X_\infty} [|A_0 - D_0|^3]$, the remaining proof for part (a) of Lemma 1 follows from (3.3).

195 For part (b) of Lemma 1, the above issue on bounding $\mathbb{E}_{X_\infty} [|A_0 - D_0|^3]$ no longer exists, because it is in the order of μN and can be bounded above by a constant under the condition $s \geq 1$. However, the inclusion of $q \in [0, 1/2)$ poses an additional challenge, where the term $1/|\zeta|$ appearing in (3.3) and (3.4) can be larger than a constant in terms of the order of magnitude; in the extreme case of $q = 0$ (NDS regime), it is in the order of $R^{1/2}$. To address this challenge, we 200 explore the property of the idle probability $\mathbb{P}(X_\infty \leq N)$ so that we can bound $1/|\zeta| \mathbb{P}(X_\infty \leq N)$ together. Then, the remaining proof for part (b) of Lemma 1 follows from (3.4).

Lemma 3 (Idle probability). *For all Λ, N , and μ satisfying $N \geq 1$, $0 < \Lambda < N\mu$, and $0 < \mu < 1$,*

$$\mathbb{P}(X_\infty \leq N) \leq \frac{1}{1 - \mu} \left[(2 + \delta)(|\zeta| + \mu\sqrt{R}) \right]. \quad (3.5)$$

We outline a coupling argument to prove this lemma, and leave the complete details of the proof in Appendix A. First, we construct a middle system that bridges our discrete queue and

an $M/M/N$ queue. This middle system is a continuous-time queue with a time-homogeneous Poisson arrival process and a *two-time-scale* service time component introduced in [1], and we use $Y^M = \{Y_k^M\}$ to denote the *midnight* customer count of this middle system. First, we show that

$$\mathbb{P}(X_\infty \leq N) = \mathbb{P}(Y_\infty^M \leq N), \quad (3.6)$$

using the fact that Y^M has the same dynamics as that of the DTMC X in our discrete queue. Then, we construct a $M/M/N$ queue such that its *midnight* customer count Y^C is always stochastically smaller than Y^M , which gives

$$\mathbb{P}(Y_\infty^M \leq N) \leq \mathbb{P}(Y_\infty^C \leq N). \quad (3.7)$$

Finally, we prove that the stationary distribution of the midnight count Y^C is the same as that of the regular customer count process, $X^C = \{X^C(t), t \geq 0\}$, in the $M/M/N$ queue, for which we can obtain an upper bound on the idle probability $\mathbb{P}(X^C(\infty) \leq N)$; see (3.15) in Lemma 2 of [19]. We eventually have (3.5) from (3.6) to (3.7).

3.2. Lemma 2

To establish the bounds on f_h''' as in (2.13), we need to first bound f' and f'' because

$$f_h'''(x) = - \left(\frac{2b(x)}{a(x)} \right)' f_h'(x) - \left(\frac{2b(x)}{a(x)} \right) f_h''(x) - \frac{2}{a(x)} h'(x) - \frac{2a'(x)}{a^2(x)} [\mathbb{E}h(Y_\infty) - h(x)], \quad (3.8)$$

where $a'(x)$ is interpreted as the left derivative at the points $x = -1/\delta$ and $x = -\zeta$. To bound $|f_h'|$, one expression for f_h' is given by

$$f_h'(x) = \frac{1}{q(x)} \int_{-\infty}^x \frac{2}{a(y)} (\mathbb{E}h(Y_\infty) - h(y)) q(y) dy, \quad x \in \mathbb{R}, \quad (3.9)$$

where

$$q(x) \equiv \exp \left(\int_0^x \frac{2b(y)}{a(y)} dy \right). \quad (3.10)$$

Correspondingly, we need to bound

$$\frac{1}{q(x)} \int_{-\infty}^x \frac{2}{a(y)} q(y) dy, \quad \frac{1}{q(x)} \int_{-\infty}^x \frac{2}{a(y)} |y| q(y) dy, \quad (3.11)$$

and $\mathbb{E}|Y_\infty|$ since $|h(x)| \leq |x|$.

If the diffusion coefficient were a constant such as the one used in Dai and Shi [1], the density $q(x)$ would have a simple Gaussian or exponential form, in which case one could directly bound (3.11)

by evaluating the integrals. However, in this paper, $a(x)$ is not only state-dependent, but also has a complicated form as given in (1.6). As a result, $q(x)$ becomes very complicated and the direct method no longer works. To overcome the difficulty, we borrow a new technique mentioned by the authors of [8] on the gradient bounds for Poisson equation with general $b(x)$ and $a(x)$. We illustrate the main idea with the first quantity in (3.11): for $x \leq -1$, since $b(x)/b(y) \geq 1$

$$\begin{aligned}
\frac{1}{q(x)} \int_{-\infty}^x \frac{2}{a(y)} q(y) dy &\leq \frac{1}{q(x)} \int_{-\infty}^x \frac{2b(y)}{a(y)} \frac{1}{b(x)} q(y) dy \\
&= \frac{1}{q(x)} \frac{1}{b(x)} \int_{-\infty}^x \frac{2b(y)}{a(y)} e^{\int_0^y \frac{2b(u)}{a(u)} du} dy \\
&= \frac{1}{q(x)} \frac{1}{b(x)} (q(x) - q(-\infty)) \\
&\leq \frac{1}{b(x)} \leq \frac{1}{\mu}.
\end{aligned} \tag{3.12}$$

There are, of course, many more integrals to deal with, which come from expressions of f'_h , f''_h and f'''_h . For each of these integrals, one may need to take different approaches for x in different ranges.

210 Besides, due to the complicated density $p(x)$, bounding $\mathbb{E}|Y_\infty|$ also needs more evolved technique. The complete proof is thus much more lengthy and is deferred to the online supplement [2].

4. Numerical results

In this section, we perform extensive numerical experiments for two purposes: (i) demonstrate the quality of using Y_∞ to approximate the stationary distribution of the DTMC X ; and (ii) provide
215 support for findings from Theorems 1 and 2. First, we choose a “baseline” experimental setting, with parameters estimated from the hospital data used in [1], i.e., $\Lambda = 90.95$, $\mu = 1/5.3$, and N around 500. Then, to demonstrate the convergence of the error bounds under different system load conditions and rates of μ converging to 0, we choose the parameter q in (1.8) to be 1, $1/2$, and 0 (corresponding to QD, QED, and NDS regime, respectively), and vary the parameter s from 0 to 3.
220 To determine the relationship between N and R in (1.8), we choose $\beta = 0.0995$ for the QD regime, $\beta = 0.9994$ for the QED regime, and $\beta = 13$ for the NDS regime so that N and R match with the baseline parameter setting.

In the experiments, we evaluate the following performance measures: the expected queue length $\mathbb{E}(X_\infty - N)^+$, its adjusted version $\mathbb{E}(X_\infty - R)^+$, and the expected number of busy servers $\mathbb{E}(X_k \wedge N)$.
225 One can easily check that the three measures can be represented by $\sqrt{R} \mathbb{E}[h(\tilde{X}_\infty)]$ with $h(x) =$

$(x + \zeta)^+$, $h(x) = x^+$, and $h(x) = \delta N - (x + \zeta)^-$, respectively, all satisfying $h(x) \in \text{Lip}(1)$. We compare each of the performance measures calculated from (i) π solved from the exact Markov chain analysis, and (ii) π approximated from using the density function p of Y_∞ given by (1.4). In the interest of space, we mainly report the expected queue length for QED and NDS regimes below.

230 We leave supporting tables on the QD regime in Appendix C. In the last column of each table, we report the *scaled error*, which equals the absolute error (between the performance calculated from (i) and (ii)) scaled by $1/\sqrt{R}$. In other words, the scaled error corresponds to the left side of (1.10) and (1.11). Also note that in a given regime, we choose the same set of N and R and change μ using different values of s .

235 Tables 1 and 2 demonstrate the results for the QED and NDS regimes, respectively, with μ being fixed at $1/5.3$. Clearly, we can see that the scaled error does not converge to 0 when we let R and N grow towards infinity. It is particularly evident for the NDS regime. Indeed, this observation has motivated us to realize the important role that μ plays in the convergence of the error bounds, and led us to identify the important asymptotic regime, i.e., to ensure the convergence of the error

240 bounds, μ also needs to converge to 0 as $N, R \rightarrow \infty$.

		$\mathbb{E}(X_\infty - N)^+$		
N	R	Stein	Exact	scaled error
504	482.06	4.78	4.57	0.93%
995	963.97	6.71	6.44	0.89%
1484	1446.00	8.20	7.85	0.90%
1972	1928.12	9.45	9.05	0.91%
2946	2892.25	11.55	11.05	0.92%
3919	3856.93	13.32	12.74	0.92%

Table 1: QED regime with $\beta = 0.9994$ and μ fixed at $1/5.3$.

		$\mathbb{E}(X_\infty - N)^+$		
N	R	Stein	Exact	scaled error
495	482	14.94	15.02	0.37%
977	964	38.23	39.00	2.46%
1459	1446	63.82	65.49	4.41%
1941	1928	90.63	93.50	6.55%
2905	2892	146.38	152.56	11.48%
3869	3856	203.91	214.63	17.27%

Table 2: NDS regime with $\beta = 13$ and μ fixed at $1/5.3$.

Next, instead of fixing μ , we let μ decrease to 0 at different rates. Table 3 and 4 report the results for the QED and the NDS regime, respectively. Table C.9 in the appendix reports the results in the QD regime. In each table, the left part corresponds to the case where $\mu \sim 1/R^{1/2}$ (i.e., $s = 1/2$), and the right part corresponds to the case where $\mu \sim 1/R$ (i.e., $s = 1$). We observe that when

245 $\mu \sim 1/R^{1/2}$ and $\mu \sim 1/R$, the scaled error converges to 0 as R, N increase under both the QD and QED regime. In contrast, Table 4 clearly shows that in the NDS regime, when $\mu \sim 1/R^{1/2}$, the scaled error does *not* converge to 0; in fact, it keeps increasing as R and N become large. When $\mu \sim 1/R$, the scaled error starts to converge to 0. These observations are also consistent with our theorems, that is, not only does μ need to converge, but it needs to converge at a fast enough rate
250 to ensure the convergence of the error bounds.

N	R	$\mathbb{E}(X_\infty - N)^+$				$\mathbb{E}(X_\infty - N)^+$			
		μ	Stein	Exact	scaled error	μ	Stein	Exact	scaled error
504	482.06	0.189	4.78	4.57	0.93%	0.189	4.78	4.57	0.93%
995	963.97	0.133	6.83	6.62	0.66%	0.094	6.90	6.76	0.48%
1484	1446.00	0.109	8.39	8.18	0.55%	0.063	8.50	8.38	0.33%
1972	1928.12	0.094	9.71	9.50	0.48%	0.047	9.84	9.73	0.25%
2946	2892.25	0.077	11.92	11.71	0.40%	0.031	12.07	11.98	0.17%
3919	3856.93	0.067	13.79	13.57	0.35%	0.024	13.95	13.87	0.13%

Table 3: QED regime with the parameter $\beta = 0.9994$. On the left, $\mu = 4.1426/R^{1/2}$ and on the right, $\mu = 90.9542/R$; in both cases the scaled error decreases towards 0, but the right side decreases much faster.

From all the results reported in this section, we can see that the approximation works remarkably well under a variety of system load conditions and assumptions on μ . In addition, even for small to moderately sized systems, the approximation still works well. Table 5 and 6 summarize the results for $N = 18$ and $N = 66$ respectively, with $\mu = 1/5.3$ and the system utilization $\rho \equiv R/N$
255 varying between 88% to 96%. Note that in these two tables we report the relative approximation error, i.e., absolute error divided by the exact value. For comparison purpose, we also display the approximation results using Y_∞^0 – the r.v. with the diffusion coefficient being a constant 2μ used in [1]. Clearly, Y_∞ significantly improves the approximation quality comparing to Y_∞^0 . For example, when $N = 18$ and $\rho = 90\%$, the relative error is only 0.51% using Y_∞ , but is 5.62% using Y_∞^0 .

N	R	$\mathbb{E}(X_\infty - N)^+$				$\mathbb{E}(X_\infty - N)^+$			
		μ	Stein	Exact	scaled error	μ	Stein	Exact	scaled error
495	482	0.189	14.94	15.02	0.37%	0.189	15.02	14.95	0.37%
977	964	0.133	39.46	39.60	0.45%	0.094	40.32	40.42	0.31%
1459	1446	0.109	66.79	67.01	0.56%	0.063	68.51	68.63	0.31%
1941	1928	0.094	95.62	95.89	0.60%	0.047	98.12	98.25	0.29%
2905	2892	0.077	155.91	156.23	0.60%	0.031	159.80	159.93	0.24%
3869	3856	0.067	218.35	218.72	0.58%	0.024	223.46	223.58	0.20%

Table 4: NDS regime with the parameter $\beta = 13$. On the left, $\mu = 4.1424/R^{1/2}$ and the scaled error increases; on the right, $\mu = 90.9434/R$ and the scaled error decreases towards 0.

ρ	$\mathbb{E}(X_\infty - N)^+$	$\sqrt{R}\mathbb{E}(Y_\infty + \zeta)^+$	relative error	$\sqrt{R}\mathbb{E}(Y_\infty^0 + \zeta)^+$	relative error
88%	3.33	3.32	0.40%	3.47	4.10%
90%	4.65	4.62	0.51%	4.91	5.62%
92%	6.67	6.68	0.02%	7.18	7.51%
94%	9.93	10.23	3.00%	11.10	11.81%
96%	15.11	17.55	16.11%	19.19	26.99%

Table 5: Approximations of the expected queue length using Y_∞ and Y_∞^0 for $N = 18$ and $\mu = 1/5.3$.

ρ	$\mathbb{E}(X_\infty - N)^+$	$\sqrt{R}\mathbb{E}(Y_\infty + \zeta)^+$	relative error	$\sqrt{R}\mathbb{E}(Y_\infty^0 + \zeta)^+$	relative error
88%	1.50	1.57	4.19%	1.53	1.46%
90%	2.48	2.53	1.72%	2.58	3.77%
92%	4.18	4.19	0.24%	4.42	5.69%
94%	7.34	7.30	0.51%	7.87	7.26%
96%	14.24	14.14	0.70%	15.45	8.54%

Table 6: Approximations of the expected queue length using Y_∞ and Y_∞^0 for $N = 66$ and $\mu = 1/5.3$.

260 5. Conclusion

In this paper, we apply the Stein’s method framework to identify a continuous random variable Y_∞ to approximate the stationary distribution of the scaled customer count, \tilde{X}_∞ , in a discrete-time queueing system. Using this framework, we characterize the convergence rate of the error bounds between \tilde{X}_∞ and Y_∞ under different system load conditions. Different from the continuous-time systems, we identify the important role of μ in the converge rate of the error bounds. The numerical results support our theoretical findings.

This work could be extended in several directions. First, under the current queueing setting, it remains to identify the accurate “cutoff” point for s that is required to ensure the convergence of the error bounds in each operation regime. Second, a major limitation of this paper is the geometric service time assumption. Following [18], one could adapt the approximation developed in this paper to queueing systems with discrete phase-type service distributions (i.e., replacing the exponential with geometric distributions in the regular phase-type distributions). This potentially could lead to efficient algorithms to analyze systems with non-geometric service time distributions. Third, since our discrete queue is motivated from the hospital inpatient flow management, a variety of model features that are important in the healthcare context can be added to the current system, for example, including the day-of-week phenomenon (which requires a discrete version of time-varying arrival process), and multiple customer classes to represent patients with different characteristics.

Acknowledgement

We thank Jim Dai and Anton Braverman for their useful comments to this paper.

280 References

- [1] J. G. Dai, P. Shi, [A two-time-scale approach to time-varying queues in hospital inpatient flow management](#), Operations Research, *forthcoming*.
URL http://papers.ssrn.com/sol3/papers.cfm?abstract_id=2489533
- [2] J. Feng, P. Shi, Online supplement to steady-state diffusion approximations for discrete-time queue in hospital inpatient flow management, online supplement (2016).
URL <https://arxiv.org/abs/1612.00790>
- [3] A. L. Gibbs, F. E. Su, [On choosing and bounding probability metrics](#), International Statistical Review / Revue Internationale de Statistique 70 (3) (2002) pp. 419–435.
URL <http://www.jstor.org/stable/1403865>
- [4] R. Atar, A diffusion regime with nondegenerate slowdown, Operations Research 60 (2) (2012) 490–500.
- [5] N. Gans, G. Koole, A. Mandelbaum, Telephone call centers: Tutorial, review, and research prospects, Manufacturing & Service Operations Management 5 (2) (2003) 79–141.

- [6] P. Shi, J. G. Dai, D. Ding, S. K. J. Ang, M. Chou, X. Jin, J. Sim, Patient flow from emergency department to inpatient wards: Empirical observations from a Singaporean hospital (2014).
URL <http://dx.doi.org/10.2139/ssrn.2517050>
- [7] National Center for Health Statistics, Health, united states, 2015: With special feature on racial and ethnic health disparities, Tech. rep., Hyattsville, Maryland (2016).
- [8] A. Braverman, J. G. Dai, High order steady-state diffusion approximation of the Erlang-C system, submitted for publication (2016).
URL <http://arxiv.org/abs/1602.02866>
- [9] A. J. E. M. Janssen, J. S. H. van Leeuwen, Back to the roots of the M/D/s queue and the works of Erlang, Crommelin and Pollaczek, Statistica Neerlandica 62 (3) (2008) 299–313. doi:10.1111/j.1467-9574.2008.00395.x.
- [10] P. Shi, M. C. Chou, J. G. Dai, D. Ding, J. Sim, Models and insights for hospital inpatient operations: Time-dependent ed boarding time, Management Science 62 (1) (2016) 1–28.
- [11] M. Armony, S. Israelit, A. Mandelbaum, Y. N. Marmor, Y. Tseytlin, G. B. Yom-Tov, Patient flow in hospitals: A data-based queueing-science perspective, Stochastic Systems 5 (1) (2015) 146–194. doi:10.1214/14-SSY153.
- [12] P. Shi, Stochastic modeling and decision making in two healthcare applications: Inpatient flow management and influenza pandemics, Ph.D. thesis, Georgia Institute of Technology (2013).
- [13] N. D. Michael Simon, Yevhen Yankovskyy, Solving the mystery of patient days and the midnight census, Nursing Management (Springhouse) (2) (2010) 1214.
- [14] I. Rubin, Z. Zhang, Message delay and queue-size analysis for circuit-switched tdma systems, IEEE Transactions on Communications 39 (6) (1991) 905–914. doi:10.1109/26.87180.
- [15] K. Laevens, H. Bruneel, Delay analysis for discrete-time queueing systems with multiple randomly interrupted servers, European Journal of Operational Research 85 (1995) 16177.
- [16] S. Chatterjee, A short survey of Stein’s method, Proceedings of ICM (2014).
URL <http://arxiv.org/abs/1404.1392>
- [17] N. Ross, Fundamentals of Stein’s method, Probab. Surv. 8 (2011) 210–293. doi:10.1214/11-PS182.
- [18] A. Braverman, J. G. Dai, Stein’s method for steady-state diffusion approximations of $M/Ph/n + M$ systems, The Annals of Applied Probability, forthcoming.
URL <http://arxiv.org/abs/1503.00774>
- [19] A. Braverman, J. G. Dai, J. Feng, Stein’s method for steady-state diffusion approximations: an introduction through the Erlang-A and B systems, submitted for publication (2015).
URL <http://arxiv.org/abs/1512.09364>
- [20] J. Huang, I. Gurvich, Beyond heavy-traffic regimes: Universal bounds and controls for the single-server queue, submitted for publication (2016).
URL http://papers.ssrn.com/sol3/papers.cfm?abstract_id=2784752
- [21] H. Bruneel, B. G. Kim, Discrete-Time Models for Communication Systems Including ATM, 1st Edition, Springer US, Boston, MA, 1993.
- [22] P. Gao, S. Wittevrongel, H. Bruneel, Discrete-time multiserver queues with geometric service times, Computers & Operations Research 31 (1) (2004) 81–99.
- [23] A. J. E. M. Janssen, J. S. H. van Leeuwen, B. Zwart, Corrected asymptotics for a multi-server queue in the Halfin-Whitt regime, Queueing Syst. 58 (4) (2008) 261–301. doi:10.1007/s11134-008-9070-0.
- [24] M. Baron, Probability and statistics for computer scientists, 2nd Edition, CRC press, 2013.
- [25] K. Sigman, Recurrence and stationary distributions, lecture note.
URL <http://www.columbia.edu/~ks20/stochastic-I/stochastic-I-MCII.pdf>
- [26] K. Sigman, Continuous-time markov chains, lecture note.
URL <http://www.columbia.edu/~ks20/stochastic-I/stochastic-I-CTMC.pdf>

340 Appendix A. Proof of Lemma 3

Proof. As mentioned in the main paper, we first construct a “middle system”, denoted as *System M*, to bridge our discrete queue and an $M/M/N$ queue. This middle system has N identical servers and a buffer of infinite size. Customers arrive according to a homogeneous Poisson process with rate Λ , and for each customer, the service time S^M follows a “two-time-scale” form [1]

$$S^M = \begin{cases} \text{LOS}^M + (1 - h_{\text{adm}}^M), & 0 < h_{\text{adm}}^M < 1, \\ \text{LOS}^M, & h_{\text{adm}}^M = 0. \end{cases} \quad (\text{A.1})$$

Here, LOS^M denotes the *number of discrete time epochs* that the customer occupies a server which takes values on $1, 2, \dots$, and we assume it follows a geometric distribution with mean $1/\mu$; h_{adm}^M is a number between 0 (inclusive) and 1 (exclusive) that denotes the instant within a discrete time epoch when the customer is admitted. Mathematically,

$$h_{\text{adm}}^M = \text{adm}^M - \lfloor \text{adm}^M \rfloor, \quad (\text{A.2})$$

where adm^M is the admission time of the customer.

Now, let $X^M = \{X^M(t) : t \geq 0\}$ denote the customer count process of system M and define its *discrete-time-epoch count* $Y^M = \{Y_k^M : k = 0, 1, \dots\}$ as

$$Y_k^M = X^M(k) = X^M(k-1) + A^M(k-1, k] - D^M(k-1, k], \quad (\text{A.3})$$

where $A^M(k-1, k]$ and $D^M(k-1, k]$ denote the total numbers of arrivals and departures occurred between time $k-1$ (exclusive) and k (inclusive), respectively.

This process Y^M is referred to as the “midnight count process” in [1], and indeed, Y^M has exactly the same dynamics as our DTMC X characterized in (1.1). Because (i) $A^M(k-1, k]$ follows a Poisson distribution with mean Λ ; and (ii) for $D^M(k-1, k]$, because of the geometric assumption on LOS^M , using the coin-toss argument in [1], we can see $D^M(k-1, k]$ follows a binomial distribution with parameters (Z_{k-1}^M, μ) , with $Z_{k-1}^M = Y_{k-1}^M \wedge N$ denoting the number of busy servers at time $k-1$. As a result, when the same stability condition $\Lambda < N\mu$ holds, the stationary distribution of Y^M uniquely exists and equals π . The corresponding steady-state random variable, Y_∞^M , satisfies

$$\mathbb{P}(Y_\infty^M \leq N) = \mathbb{P}(X_\infty \leq N). \quad (\text{A.4})$$

Next, we consider the $M/M/N$ queue (Erlang-C model) where customers arrive according to a homogeneous Poisson process at rate Λ , and the service time for each customer, S^C , has an exponential distribution with rate $\mu^C = -\log(1-\mu) > 0$.

Let $X^C = \{X^C(t) : t \geq 0\}$ denote the customer count process of this Erlang-C system, and define its discrete-time-epoch count $Y^C = \{Y_k^C : k = 0, 1, \dots\}$ as

$$Y_k^C = X^C(k) = X^C(k-1) + A^C(k-1, k] - D^C(k-1, k], \quad (\text{A.5})$$

where $A^C(k-1, k]$ and $D^C(k-1, k]$ are the total numbers of arrivals and departures occurred between time $k-1$ (exclusive) and k (inclusive), respectively, in this Erlang-C system.

Next, we use a *coupling argument* to show that, on any given sample path, the discrete-time-epoch count of this Erlang-C system is always less than that of System M , which gives

$$\mathbb{P}(Y_\infty^M \leq N) \leq \mathbb{P}(Y_\infty^C \leq N). \quad (\text{A.6})$$

To do so, for a given sample path, we construct a stream of customers, with index $i = 1, 2, \dots$ to arrive to the Erlang-C system at time t_1, t_2, \dots . These customers are pre-designated with service times s_1, s_2, \dots , sampled from the exponential distribution with rate μ^C . Denote these customers’ admission and departure times as $\text{adm}_1^C, \text{adm}_2^C, \dots$, and $\text{dis}_1^C, \text{dis}_2^C, \dots$, respectively.

To couple system M with the Erlang-C system, we construct another stream of customers to arrive to System M , also with index $i = 1, 2, \dots$ and arrive at the exact same time t_1, t_2, \dots . Let their LOS^M be $\lceil s_1 \rceil, \lceil s_2 \rceil, \dots$, and their service times be calculated from (A.1). According to Section 5.2.3 of [24], $\lceil S^C \rceil$ follows a geometric distribution with success probability $1 - \exp(-\mu^C)$, which exactly equals μ . Hence, $\lceil s_1 \rceil, \lceil s_2 \rceil, \dots$ are indeed generated from a geometric distribution whose mean is $1/\mu$ (and takes values on $1, 2, \dots$). Denote these customers’ admission and departure times in System M as $\text{adm}_1^M, \text{adm}_2^M, \dots$, and $\text{dis}_1^M, \text{dis}_2^M, \dots$, respectively.

The following lemma shows that for each customer, the admission and departure times in System M are always earlier than those in the Erlang-C system.

Lemma 4. *On any given sample path, for each customer $i = 1, 2, \dots$,*

$$\text{adm}_i^C \leq \text{adm}_i^M, \quad \text{dis}_i^C \leq \text{dis}_i^M. \quad (\text{A.7})$$

The proof is given in [Appendix A.1](#). We know that for each i , customer i arrives to both systems at time t_i , and departs at time dis_i^C from the Erlang-C system. Then, this customer is included in Y^C at and only at the discrete time epochs $\lceil t_i \rceil, \lceil t_i \rceil + 1, \dots, \lceil \text{dis}_i^C \rceil - 1$. By Lemma 4, $\text{dis}_i^C \leq \text{dis}_i^M$. This implies that the customer is included in Y^M at least at the discrete time epochs $\lceil t_i \rceil, \lceil t_i \rceil + 1, \dots, \lceil \text{dis}_i^C \rceil - 1$. Since this observation is true for every customer, we conclude that $Y_k^C \leq Y_k^M$ for all $k = 0, 1, \dots$, which proves (A.6).

For the Erlang-C system, one can show that the discrete count process Y^C is an irreducible Markov chain which is positive recurrent with a unique stationary distribution, π^Y , under the condition that

$$\Lambda/\mu^C < N, \quad (\text{A.8})$$

which always holds when (1.2) is satisfied due to the inequality $\mu < -\log(1 - \mu)$. Then, from (A.6) and (A.4),

$$\mathbb{P}(X_\infty \leq N) \leq \mathbb{P}(Y_\infty^C \leq N), \quad (\text{A.9})$$

where Y_∞^C denotes the steady-state random variable of Y^C .

Finally, we show that in the Erlang-C queue,

$$\mathbb{P}(Y_\infty^C \leq N) = \mathbb{P}(X^C(\infty) \leq N). \quad (\text{A.10})$$

To do so, note that Y^C is aperiodic. Thus, we have the following relationship according to Proposition 2.9 of [25]

$$\pi_n^Y = \lim_{k \rightarrow \infty} \mathbb{P}(Y_k^C = n | Y_0^C = m), \quad m, n \in \mathbb{N}, \quad (\text{A.11})$$

where the type of convergence above is weak convergence. Also note that (A.8) guarantees that X^C is positive recurrent with a unique stationary distribution. Denote this distribution with π^X , and the corresponding random variable with $X^C(\infty)$. According to Proposition 1.1 of [26], we have

$$\pi_n^X = \lim_{t \rightarrow \infty} \mathbb{P}(X^C(t) = n | X^C(0) = m), \quad m, n \in \mathbb{N}, \quad (\text{A.12})$$

where the type of convergence above is weak convergence.

Combining (A.11) and (A.12), it is immediately seen that the stationary distributions of X^C and Y^C are the same, which implies (A.10). Applying Lemma 2 of [19] to the right side of (A.10), we obtain

$$\begin{aligned} \mathbb{P}(X^C(\infty) \leq N) &\leq \left(2 + \frac{1}{\sqrt{\Lambda/\mu^C}}\right) \frac{1}{\sqrt{\Lambda/\mu^C}} (N - \Lambda/\mu^C) \\ &\leq \left(2 + \frac{1}{\sqrt{\Lambda/\nu}}\right) \frac{1}{\sqrt{\Lambda/\nu}} (N - \Lambda/\nu) \\ &= 2 \frac{1}{\sqrt{R(1-\mu)}} [N - R(1-\mu)] + \frac{1}{R(1-\mu)} [N - R(1-\mu)] \\ &= \frac{2}{\sqrt{1-\mu}} (|\zeta| + \mu\sqrt{R}) + \frac{1}{1-\mu} (\delta|\zeta| + \mu) \\ &\leq \frac{1}{1-\mu} (2 + \delta)(|\zeta| + \mu\sqrt{R}), \end{aligned} \quad (\text{A.13})$$

where $\nu = \mu/(1 - \mu)$, and the second inequality comes from the fact that for $\mu \in (0, 1)$, $-\log(1 - \mu) < \mu/(1 - \mu)$.

Combining (A.9), (A.10), and (A.13) establishes Lemma 3. \square

370 Appendix A.1. Proof of Lemma 4

We prove (A.7) by induction, starting with $i = 1$.

For customer 1 in both systems, she arrives at time t_1 and is admitted immediately. That is,

$$\text{adm}_1^C = \text{adm}_1^M = t_1. \quad (\text{A.14})$$

For the departure time of this customer, in the Erlang-C system,

$$\text{dis}_1^C = \text{adm}_1^C + s_1 = t_1 + s_1. \quad (\text{A.15})$$

In System M , LOS^M for this customer is $\lceil s_1 \rceil$. Since the service time S^M is always greater than or equal to LOS^M , we have

$$\text{dis}_1^M \geq \text{adm}_1^M + \lceil s_1 \rceil \geq t_1 + s_1. \quad (\text{A.16})$$

Together, (A.14), (A.15), and (A.16) proves (A.7) for $i = 1$.

Now, assume (A.7) holds for the first i customers, where $1 \leq i \in \mathbb{N}$. Consider the next customer, $i + 1$. We claim that the admission time for this customer satisfies

$$t_{i+1} \leq \text{adm}_{i+1}^C \leq \text{adm}_{i+1}^M. \quad (\text{A.17})$$

Suppose the otherwise. Then at time adm_{i+1}^M , N different customers, $j_1, j_2, \dots, j_N \leq i$, are being served by the N servers in the Erlang-C system, whereas at least one of them, j^* , has departed from System M . This leads to a contradiction, since we are assuming that (A.7) holds for all $j \leq i$, and thus in particular for j^* .

With (A.17) in hand, the departure time of customer $i + 1$ satisfies

$$\text{dis}_{i+1}^C = \text{adm}_{i+1}^C + s_{i+1} \leq \text{adm}_{i+1}^M + \lceil s_{i+1} \rceil \leq \text{dis}_{i+1}^M. \quad (\text{A.18})$$

Combining (A.17) and (A.18), we have shown that (A.7) holds for the first $i + 1$ customers. This finishes the induction step and proves Lemma 4.

Appendix B. Computational time to evaluate π using the Markov chain analysis

All the experiments are implemented in Matlab and run on a Linux 64-bit cluster hosted either by a Dell R620 server or by a R720 server (indicated by an “*”) with 1 processor. The virtual size supplied by the cluster is 8192 MiB for each experiment with $N = 4242$ or 977.

The computational time of these experiments is shown in Table B.7.

μ	elapsed time (day)	μ	elapsed time (hour)
0.189	1.25	0.189	1.34
0.067	0.84*	0.133	0.96*
0.040	0.84*	0.112	0.93*
0.024	0.87	0.094	1.18*
0.008	1.13	0.067	0.97*

$N = 4242, \rho = 90\%$

$N = 977, \rho = 98\%$

Table B.7: Computational time of π for large systems with moderately high utilization (left) and for moderately large systems with high utilization (right).

Appendix C. Numerical results in the QD regime

This section includes numerical results for the QD regime. In this regime, the expected queue length is very close to 0 because of the light system load. Hence we report an adjusted version of the queue length, $\mathbb{E}(X_\infty - R)^+$, in Table C.8 and C.9.

N	R	$\mathbb{E}(X_\infty - R)^+$		
		Stein	Exact	scaled error
530	482.06	8.80	8.91	0.48%
1061	965.02	12.31	12.40	0.87%
1591	1447.08	14.80	15.18	0.98%
2121	1929.14	17.08	17.52	1.01%
3182	2894.16	20.91	21.50	1.10%
4242	3858.28	24.15	24.78	1.02%

Table C.8: QD regime with the parameter $\beta = 0.0995$ and μ fixed at $1/5.3$.

N	R	μ	$\mathbb{E}(X_\infty - R)^+$			μ	$\mathbb{E}(X_\infty - R)^+$		
			Stein	Exact	scaled error		Stein	Exact	scaled error
530	482.06	0.189	8.80	8.91	0.48%	0.189	8.80	8.91	0.48%
1061	965.02	0.133	12.21	12.40	0.61%	0.094	12.27	12.40	0.43%
1591	1447.08	0.109	14.97	15.18	0.55%	0.063	15.06	15.18	0.32%
2121	1929.14	0.094	17.31	17.52	0.49%	0.047	17.42	17.52	0.24%
3182	2894.16	0.077	21.25	21.46	0.40%	0.031	21.38	21.46	0.16%
4242	3858.28	0.067	24.57	24.78	0.34%	0.024	24.71	24.78	0.12%

Table C.9: QD regime with the parameter $\beta = 0.0995$. On the left, $\mu = 4.1426/R^{1/2}$ and on the right, $\mu = 90.9542/R$; in both cases the scaled error decreases towards 0 at a very similar rate.

Online Supplement of “Steady-state Diffusion Approximations for Discrete-time Queue in Hospital Inpatient Flow Management”

Jiekun Feng

Department of Statistical Science, Cornell University

Pengyi Shi

Krannert School of Management, Purdue University

This document serves as the online supplement for Feng and Shi [1], which we refer to as the “main paper.” In the main paper, we analyze a $GI/Geo/N$ discrete-time queue (or *discrete queue* in short), and use the Stein’s method framework to develop steady-state diffusion approximations for the customer count process, with a focus on the Poisson arrival case. We establish the error

5 bounds of the approximations in Theorems 1 and 2 there. The proof of these two theorems rely on Lemmas 1 and 2, for which we give out the complete proof in Sections 1 and 4 of this document, respectively. The proof of Lemma 1 further depends on several additional lemmas, which we prove in Section 2. Section 3 of this document extends the results in the main paper by considering a

10 general arrival distribution, where we develop analogous steady-state approximations and establish the corresponding error bounds.

Email addresses: jf646@cornell.edu (Jiekun Feng), shi178@purdue.edu (Pengyi Shi)

1. Proof of Lemma 1 in Feng and Shi [1]

Let $f = f_h$ be a solution to the Poisson equation defined in (2.2) of the main paper. To prove this lemma, we first consider $\epsilon(X_\infty)$ when $X_\infty \leq N$, or equivalently, $\tilde{X}_\infty \leq -\zeta$,

$$\begin{aligned}
\epsilon(X_\infty) &= \mathbb{E}_{X_\infty} \left[\left| \int_{\tilde{X}_\infty}^{\tilde{X}_\infty + \delta(A_0 - D_0)} |f'''(y)| dy \right| (A_0 - D_0)^2 \mathbf{1}_{\{X_\infty + A_0 - D_0 \leq N\}} \right] \\
&\quad + \mathbb{E}_{X_\infty} \left[\left| \int_{\tilde{X}_\infty}^{\tilde{X}_\infty + \delta(A_0 - D_0)} |f'''(y)| dy \right| (A_0 - D_0)^2 \mathbf{1}_{\{X_\infty + A_0 - D_0 > N\}} \right] \\
&\leq \mathbb{E}_{X_\infty} \left[\left| \int_{\tilde{X}_\infty}^{\tilde{X}_\infty + \delta(A_0 - D_0)} \frac{C_0}{\mu} (1 + 1/|\zeta|) dy \right| (A_0 - D_0)^2 \mathbf{1}_{\{X_\infty + A_0 - D_0 \leq N\}} \right] \\
&\quad + \mathbb{E}_{X_\infty} \left[\left| \int_{\tilde{X}_\infty}^{-\zeta} \frac{C_0}{\mu} (1 + 1/|\zeta|) dy \right| (A_0 - D_0)^2 \mathbf{1}_{\{X_\infty + A_0 - D_0 > N\}} \right] \\
&\quad + \mathbb{E}_{X_\infty} \left[\left| \int_{-\zeta}^{\tilde{X}_\infty + \delta(A_0 - D_0)} \frac{4}{\mu} dy \right| (A_0 - D_0)^2 \mathbf{1}_{\{X_\infty + A_0 - D_0 > N\}} \right] \\
&= \delta \frac{C_0}{\mu} (1 + 1/|\zeta|) \mathbb{E}_{X_\infty} \left[|(X_\infty + A_0 - D_0) \wedge N - X_\infty| (A_0 - D_0)^2 \right] \tag{1.1} \\
&\quad + \delta \frac{4}{\mu} \mathbb{E}_{X_\infty} \left[(A_0 - D_0 - (N - X_\infty)) (A_0 - D_0)^2 \mathbf{1}_{\{X_\infty + A_0 - D_0 > N\}} \right]. \tag{1.2}
\end{aligned}$$

To get the first inequality above, we use the gradient bound for f''' proved in Lemma 2 of the main paper. The reason we need to separately consider the two cases $X_\infty + A_0 - D_0 > N$ and $X_\infty + A_0 - D_0 \leq N$ is that the gradient bounds for f''' are different on $(-\infty, -\zeta]$ and $(-\zeta, \infty)$.

Similarly, when $X_\infty > N$, or equivalently, $\tilde{X}_\infty > -\zeta$, we get

$$\epsilon(X_\infty) \leq \delta \frac{4}{\mu} \mathbb{E}_{X_\infty} \left[|X_\infty - (X_\infty + A_0 - D_0) \vee N| (A_0 - D_0)^2 \right] \tag{1.3}$$

$$+ \delta \frac{C_0}{\mu} (1 + 1/|\zeta|) \mathbb{E}_{X_\infty} \left[(D_0 - A_0 - (X_\infty - N)) (A_0 - D_0)^2 \mathbf{1}_{\{X_\infty + A_0 - D_0 \leq N\}} \right]. \tag{1.4}$$

It is easy to see that the term inside each of the expectations in (1.1) to (1.4) can be bounded above by $|A_0 - D_0|^3$. Therefore, (1.1) and (1.2) imply that

$$\begin{aligned}
&\frac{1}{2} \delta^2 \mathbb{E} [\epsilon(X_\infty) \mathbf{1}_{\{X_\infty \leq N\}}] \\
&\leq \frac{1}{2} \delta^2 \mathbb{E} \left\{ \delta \frac{C_0}{\mu} (1 + 1/|\zeta|) \mathbb{E}_{X_\infty} |A_0 - D_0|^3 \mathbf{1}_{\{X_\infty \leq N\}} \right\} + \frac{1}{2} \delta^2 \delta \frac{4}{\mu} \mathbb{E} [\mathbb{E}_{X_\infty} |A_0 - D_0|^3] \\
&\leq \delta^3 \frac{2}{\mu} \mathbb{E} [\mathbb{E}_{X_\infty} |A_0 - D_0|^3] + \frac{C_0}{2} \delta^3 \frac{1}{\mu} (1 + 1/|\zeta|) \mathbb{E} [\mathbb{E}_{X_\infty} |A_0 - D_0|^3 \mathbf{1}_{\{X_\infty \leq N\}}]. \tag{1.5}
\end{aligned}$$

Meanwhile, note that (1.3) is subject to the same upper bound as that for (1.2), which is the first term in (1.5). Thus, (1.3) and (1.4) imply that

$$\begin{aligned} & \frac{1}{2}\delta^2\mathbb{E}[\epsilon(X_\infty)\mathbb{1}_{\{X_\infty>N\}}] \\ & \leq \delta^3\frac{2}{\mu}\mathbb{E}[\mathbb{E}_{X_\infty}|A_0-D_0|^3] + \frac{C_0}{2}\delta^3\frac{1}{\mu}(1+1/|\zeta|)\mathbb{E}\left\{\mathbb{E}_{X_\infty}\left[|A_0-D_0|^3\mathbb{1}_{\{X_\infty+A_0-D_0\leq N\}}\right]\mathbb{1}_{\{X_\infty>N\}}\right\} \end{aligned} \quad (1.6)$$

$$= \delta^3\frac{2}{\mu}\mathbb{E}[\mathbb{E}_{X_\infty}|A_0-D_0|^3] + \frac{C_0}{2}\delta^3\frac{1}{\mu}(1+1/|\zeta|)\mathbb{E}\left\{|A-D|^3\mathbb{E}_{X_\infty}[\mathbb{1}_{\{X_\infty+A_0-D_0\leq N\}}]\mathbb{1}_{\{X_\infty>N\}}\right\} \quad (1.7)$$

$$\begin{aligned} & \leq \delta^3\frac{2}{\mu}\mathbb{E}[\mathbb{E}_{X_\infty}|A_0-D_0|^3] + \frac{C_0}{2}\delta^3\frac{1}{\mu}(1+1/|\zeta|)\mathbb{E}\left[|A-D|^3\right]\mathbb{P}(X_\infty+A_0-D_0\leq N) \\ & = \delta^3\frac{2}{\mu}\mathbb{E}[\mathbb{E}_{X_\infty}|A_0-D_0|^3] + \frac{C_0}{2}\delta^3\frac{1}{\mu}(1+1/|\zeta|)\mathbb{E}\left[|A-D|^3\right]\mathbb{P}(X_\infty\leq N). \end{aligned} \quad (1.8)$$

For the equality (1.7), we use the fact that when $X_\infty > N$, the number of discharges is $Bino(N, \mu)$, which is independent of X_∞ and $X_\infty + A_0 - D_0$. We adopt the two new notations for the arrival and discharge quantities, $A \sim Poiss(\Lambda)$ and $D \sim Bino(N, \mu)$, to emphasize their independence on X_∞ . Consequently, we can decouple $|A-D|^3$ from \mathbb{E}_{X_∞} . For the last equality (1.8), we use the fact that the system is in the steady-state so that the next period customer count $X_\infty + A_0 - D_0$ has the same distribution as the current count X_∞ .

As mentioned in the main paper, we address two different challenges associated with the two parts of Lemma 1, that is, one is on bounding $\mathbb{E}_{X_\infty}[|A_0 - D_0|^3]$, and the other is on bounding the idle probability $\mathbb{P}(X_\infty \leq N)$. We specify the details in the following two subsections, respectively.

1.1. Part (a) of Lemma 1

To bound $\mathbb{E}_{X_\infty}[|A_0 - D_0|^3]$, we use the c_r -inequality and Lemma 2.2 of this online supplement [2]. That is, for any $n \in \mathbb{N}$,

$$\begin{aligned} \mathbb{E}_n|A_0 - D_0|^3 &= \mathbb{E}_n[|(A_0 - \Lambda) + (z(n)\mu - D_0) + (\Lambda - z(n)\mu)|^3] \\ &\leq 9\mathbb{E}_n|A_0 - \Lambda|^3 + 9\mathbb{E}_n|D_0 - z(n)\mu|^3 + 9\mathbb{E}_n|\Lambda - z(n)\mu|^3 \\ &\leq 27\max\left(\Lambda, \Lambda^{3/2}\right) + 27\max\left(z(n)\mu, (z(n)\mu)^{3/2}\right) + 9|\Lambda - z(n)\mu|^3, \end{aligned}$$

where $z(n) = n \wedge N$.

Letting $X_\infty = n$ above and taking expectation with respect to X_∞ , we obtain

$$\mathbb{E} \left[\mathbb{E}_{X_\infty} |A_0 - D_0|^3 \right] < 54 \max \left(N\mu, (N\mu)^{3/2} \right) + 9 \mathbb{E} |z(X_\infty)\mu - \Lambda|^3, \quad (1.9)$$

where the first inequality follows from the stability condition $\Lambda < N\mu$.

Let us restate the characterization of μ and N in this supplement, i.e.,

$$\mu = \gamma R^{-s} = \gamma \delta^{2s}, \quad N = R + \beta R^q. \quad (1.10)$$

When $s \in [1/2, 1]$ and $q \in [1/2, 1]$, by (2.26), we have

$$54 \max \left(N\mu, (N\mu)^{3/2} \right) \leq 54\gamma(1 + \beta)[1 \vee \sqrt{\gamma(1 + \beta)}]\delta^{3(s-1)}. \quad (1.11)$$

The main challenge here is to bound the second term on the right side of (1.9), which is closely
30 related with the following lemma.

Lemma 1.1 (Partial third moment bound). *Consider the DTMC X and the scaled version \tilde{X} . For all $N \geq 1$, $\Lambda > 0$, and $\mu \in (0, 1)$ that satisfy $1 \leq R < N$ and (1.10) for some $s \in [1/2, 1]$ and $q \in [1/2, 1]$,*

$$\mathbb{E} \left[|\tilde{X}_\infty|^3 \mathbb{1}_{\{\tilde{X}_\infty \leq -\zeta\}} \right] \leq C_8(\gamma, \beta) \delta^{\min\{3(s-1+q/2), 0\}} \delta^{-3/4}. \quad (1.12)$$

The proof of this lemma is given in Section 2.3 of this online supplement. One can also recover the constant $C_8(\gamma, \beta)$ there. Now we are ready to bound the second term on the right side of (1.9).

Applying the moment bounds (1.12), (2.14), and the characterization of $|\zeta|$ in terms of δ as given by (2.24), we have

$$\begin{aligned} \mathbb{E} |z(X_\infty)\mu - \Lambda|^3 &= \mathbb{E} |N\mu - \Lambda - \mu(X_\infty - N)|^3 = \delta^{-3} \mathbb{E} |b(\tilde{X}_\infty)|^3 \\ &= \delta^{-3} \mu^3 \mathbb{E} [|\tilde{X}_\infty^3 \mathbb{1}_{\{\tilde{X}_\infty < -\zeta\}}|] + \delta^{-3} \mu^3 |\zeta|^3 \mathbb{P}(\tilde{X}_\infty \geq -\zeta) \\ &\leq \delta^{-3} \mu^3 C_8(\gamma, \beta) \delta^{\min\{3(s-1+q/2), 0\}} \delta^{-3/4} + \delta^{-3} \mu^3 |\zeta|^2 \left[(\delta^2 + 2) \frac{1}{|\zeta|} + \delta \right] \\ &= C_8(\gamma, \beta) \gamma^3 \delta^{\min\{3(s-1+q/2), 0\}} \delta^{6s-15/4} + \gamma^3 \beta^2 (1 + 3/\beta) \delta^{6s-2q-2}. \end{aligned} \quad (1.13)$$

Combining (1.5) and (1.6), we can establish the inequality in part (a) of Lemma 1 as follows.

$$\begin{aligned} \frac{1}{2} \delta^2 \mathbb{E} [\epsilon(X_\infty)] &\leq 4\delta^3 \frac{1}{\mu} \mathbb{E} (\mathbb{E}_{X_\infty} |A_0 - D_0|^3) + C_0 \delta^3 \frac{1}{\mu} (1 + 1/|\zeta|) \mathbb{E} (\mathbb{E}_{X_\infty} |A_0 - D_0|^3) \\ &\leq (4 + C_0) \delta^3 \frac{1}{\mu} (1 + 1/|\zeta|) \mathbb{E} (\mathbb{E}_{X_\infty} |A_0 - D_0|^3). \end{aligned} \quad (1.14)$$

Substituting (1.11) and (1.13) into (1.9), and again using the form of $|\zeta|$ in (2.24), we get from (1.14) that

$$\begin{aligned}
\frac{1}{2}\delta^2\mathbb{E}[\epsilon(X_\infty)] &\leq (4+C_0)\delta^2\delta\frac{1}{\mu}(1+1/|\zeta|)\mathbb{E}(|\mathbb{E}_{X_\infty}|A_0-D_0|^3) \\
&\leq (4+C_0)\frac{1}{\gamma}(1+1/\beta)\delta^{3-2s}\left\{54\gamma(1+\beta)[1\vee\sqrt{\gamma(1+\beta)}]\delta^{3(s-1)}\right. \\
&\quad \left.+9C_8(\gamma,\beta)\gamma^3\delta^{\min\{3(s-1+q/2),0\}}\delta^{6s-15/4}+9\gamma^3\beta^2(1+3/\beta)\delta^{6s-2q-2}\right\} \\
&\leq (4+C_0)\frac{1}{\gamma}(1+1/\beta)\left\{54\gamma(1+\beta)[1\vee\sqrt{\gamma(1+\beta)}]\delta^s\right. \\
&\quad \left.+9C_8(\gamma,\beta)\gamma^3\delta^{\min\{3(s-1+q/2),0\}}\delta^{4s-3/4}+9\gamma^3\beta^2(1+3/\beta)\delta^{4s-2q+1}\right\}. \tag{1.15}
\end{aligned}$$

Since $4s-2q+1\geq s$, $4s-3/4\geq s$, and

$$3(s-1+q/2)+4s-3/4\geq 3(s-3/4)+4s-3/4=7s-3\geq s,$$

(1.15) leads to the first half of Lemma 1 in the main paper.

1.2. Part (b) of Lemma 1

For part (b) of Lemma 1, the above issue on bounding $\mathbb{E}_{X_\infty}[|A_0-D_0|^3]$ no longer exists, because it is in the order of μN and the latter can be bounded above by a constant under the condition $s\geq 1$. To see this, recall that $\mathbb{E}_{X_\infty}|A_0-D_0|^3$ is a random variable taking on the value $\mathbb{E}_n|A_0-D_0|^3$ when $X_\infty=n$. From Lemma 2.1 of this online supplement, for any $n\in\mathbb{N}$, we have

$$\mathbb{E}_n|A_0-D_0|^3\leq 40[1\vee(N\mu)^2]N\mu.$$

Since the upper bound on the right side does not depend on n , it is an upper bound for the random variable $\mathbb{E}_{X_\infty}|A_0-D_0|^3$. Thus,

$$\mathbb{E}\left[\mathbb{E}_{X_\infty}|A_0-D_0|^3\right]\leq\mathbb{E}\left\{40[1\vee(N\mu)^2]N\mu\right\}=40[1\vee(N\mu)^2]N\mu, \tag{1.16}$$

$$\mathbb{E}\left[\mathbb{E}_{X_\infty}|A_0-D_0|^3\mathbf{1}_{\{X_\infty\leq N\}}\right]\leq\mathbb{E}\left\{40[1\vee(N\mu)^2]N\mu\mathbf{1}_{\{X_\infty\leq N\}}\right\}=40[1\vee(N\mu)^2]N\mu\mathbb{P}(X_\infty\leq N). \tag{1.17}$$

Applying Lemma 3 in the main paper on the idle probability $\mathbb{P}(X_\infty\leq N)$, we have that

$$\begin{aligned}
(1+1/|\zeta|)\mathbb{P}(X_\infty\leq N) &\leq 1+(1/|\zeta|)\mathbb{P}(X_\infty\leq N) \\
&\leq \begin{cases} 1+1/\beta, & q\geq 1/2, \\ 1+\frac{2+\delta}{1-\mu}(1+\frac{\gamma}{\beta}R^{1-s-q}), & q<1/2. \end{cases} \tag{1.18}
\end{aligned}$$

To finish proving part (b), we use

$$\begin{aligned} \frac{1}{2}\delta^2\mathbb{E}[\epsilon(X_\infty)] &\leq 4\delta^3\frac{1}{\mu}\{40[1\vee(N\mu)^2]N\mu\} + C_0\delta^3\frac{1}{\mu}\{40[1\vee(N\mu)^2]N\mu\}[(1+1/|\zeta|)\mathbb{P}(X_\infty\leq N)] \\ &= 4\delta(N\delta^2)\{40[1\vee(N\mu)^2]\}\left\{1+\frac{C_0}{4}[(1+1/|\zeta|)\mathbb{P}(X_\infty\leq N)]\right\}, \end{aligned} \quad (1.19)$$

35 where the first inequality comes from applying (1.16) and (1.17) to (1.5) and (1.8). Then, part (b) of Lemma 1 follows from (1.19) and some algebra involving (1.18), the characterization of μ , N in terms of R in (1.10), and (2.24)-(2.27) derived in the next section of this supplement.

2. Additional lemmas

2.1. Moments of random variables

Lemma 2.1 (Random variable moments). *Let $A \sim \text{Poisson}(\lambda)$, and $D \sim \text{Binomial}(M, r)$. Then*

$$\mathbb{E}A = \lambda, \mathbb{E}A^2 = \lambda + \lambda^2, \mathbb{E}A^3 = \lambda + 3\lambda^2 + \lambda^3, \mathbb{E}A^4 = \lambda + 7\lambda^2 + 6\lambda^3 + \lambda^4. \quad (2.1)$$

$$\begin{aligned} \mathbb{E}D &= Mr, \mathbb{E}D^2 = Mr(1-r+Mr), \mathbb{E}D^3 = Mr(1-3r+3Mr+2r^2-3Mr^2+M^2r^2), \\ \mathbb{E}D^4 &= Mr(1-7r+7Mr+12r^2-18Mr^2+6M^2r^2-6r^3+11Mr^3-6M^2r^3+M^3r^3). \end{aligned} \quad (2.2)$$

40 *Proof.* See [3] and [4]. □

Remark 1. Note that for $r \in [0, 1]$,

$$\begin{aligned} 1-3r+2r^2 &\leq 1, \\ 1-7r+12r^2-6r^3 &\leq 1-7r+12r^2 \leq 6, \\ 7Mr-18Mr^2+11Mr^3 &= 7Mr-Mr^2(18-11r) \leq 7Mr. \end{aligned}$$

Hence, (2.2) implies

$$\mathbb{E}D^3 \leq 5\max\{Mr, (Mr)^3\}, \quad \mathbb{E}D^4 \leq 20\max\{Mr, (Mr)^4\}. \quad (2.3)$$

Remark 2. Applying the c_r -inequality and Lemma 2.1, we have the following inequalities

$$\mathbb{E}|A-D|^3 \leq 4\mathbb{E}A^3 + 4\mathbb{E}D^3 \leq 20\max(\lambda, \lambda^3) + 20\max(Mr, (Mr)^3). \quad (2.4)$$

and

$$\mathbb{E}_n|A_0 - D_0| \leq \mathbb{E}A_0 + \mathbb{E}_nD_0 \leq 2N\mu, \quad (2.5)$$

$$\mathbb{E}_n(A_0 - D_0)^2 \leq 2\mathbb{E}A_0^2 + 2\mathbb{E}_nD_0^2 \leq 8[1 \vee (N\mu)]N\mu, \quad (2.6)$$

$$\mathbb{E}_n|A_0 - D_0|^3 \leq 4\mathbb{E}A_0^3 + 4\mathbb{E}_nD_0^3 \leq 40[1 \vee (N\mu)^2]N\mu, \quad (2.7)$$

$$\mathbb{E}_n(A_0 - D_0)^4 \leq 8\mathbb{E}A_0^4 + 8\mathbb{E}_nD_0^4 \leq 280[1 \vee (N\mu)^3]N\mu. \quad (2.8)$$

where \mathbb{E}_n is the expectation under \mathbb{P}_n , the conditional probability distribution given that the starting customer count equals n .

Lemma 2.2 (Random variable absolute central moments). *Let $A \sim \text{Poisson}(\lambda)$, and $D \sim \text{Binomial}(M, r)$. Then*

$$\mathbb{E}[|A - \lambda|^3] \leq 3 \left[\lambda \mathbb{1}_{\{\lambda < 1\}} + \lambda^{3/2} \mathbb{1}_{\{\lambda \geq 1\}} \right] = 3 \max \left(\lambda, \lambda^{3/2} \right), \quad (2.9)$$

$$\mathbb{E}[|D - Mr|^3] \leq 3 \left[Mr \mathbb{1}_{\{Mr < 1\}} + (Mr)^{3/2} \mathbb{1}_{\{Mr \geq 1\}} \right] = 3 \max \left(Mr, (Mr)^{3/2} \right). \quad (2.10)$$

Proof. From [3], the central moments of A are

$$\mathbb{E}[(A - \lambda)^4] = \lambda + 3\lambda^2, \quad \mathbb{E}[(A - \lambda)^3] = \lambda.$$

When $\lambda \geq 1$, $\mathbb{E}[(A - \lambda)^4] \leq 4\lambda^2$, and Jensen's inequality implies that

$$\mathbb{E}[|A - \lambda|^3] \leq \left\{ \mathbb{E}[(A - \lambda)^4] \right\}^{3/4} \leq 3\lambda^{3/2}.$$

When $0 < \lambda < 1$,

$$\mathbb{E}[|A - \lambda|^3] = \mathbb{E}[(A - \lambda)^3] + 2\lambda^3 \mathbb{P}(A = 0) = \lambda + 2\lambda^3 e^{-\lambda} \leq 3\lambda.$$

From [4], the central moments of D are

$$\mathbb{E}[(D - Mr)^4] = 3M^2r^2(1-r)^2 + Mr(1-r)[1 - 6r(1-r)], \quad \mathbb{E}[(D - Mr)^3] = Mr(1-r)(1-2r).$$

When $Mr \geq 1$,

$$\mathbb{E}[(D - Mr)^4] \leq 3M^2r^2 + Mr \leq 4(Mr)^2.$$

Jensen's inequality implies

$$\mathbb{E}[|D - Mr|^3] \leq \left\{ \mathbb{E}[(D - Mr)^4] \right\}^{3/4} \leq 3(Mr)^{3/2}.$$

When $0 < Mr < 1$,

$$\mathbb{E} \left[|D - Mr|^3 \right] = \mathbb{E} \left[(D - Mr)^3 \right] + 2(Mr)^3 \mathbb{P}(D = 0) \leq Mr + 2(Mr)^3 \leq 3Mr.$$

□

2.2. Moment bounds of \tilde{X}_∞

Lemma 2.3 (Moment bounds). *For all Λ, N , and μ satisfying $N \geq 1$, $0 < \Lambda < N\mu$, and $0 < \mu < 1$,*

$$\mathbb{E} \left[(\tilde{X}_\infty)^2 \mathbb{1}_{\{\tilde{X}_\infty \leq -\zeta\}} \right] \leq \frac{4}{3} + \frac{8}{3} \delta^2, \quad (2.11)$$

$$\mathbb{E} \left[|\tilde{X}_\infty \mathbb{1}_{\{\tilde{X}_\infty \leq -\zeta\}}| \right] \leq \sqrt{\frac{4}{3} + \frac{8}{3} \delta^2}, \quad (2.12)$$

$$\mathbb{E} \left[|\tilde{X}_\infty \mathbb{1}_{\{\tilde{X}_\infty \leq -\zeta\}}| \right] \leq 2|\zeta| \quad (2.13)$$

$$\mathbb{E} \left[|\tilde{X}_\infty \mathbb{1}_{\{\tilde{X}_\infty \geq -\zeta\}}| \right] \leq (\delta^2 + 1) \frac{1}{|\zeta|} + \delta, \quad (2.14)$$

$$\mathbb{E} \left[|(\tilde{X}_\infty + \zeta) \mathbb{1}_{\{\tilde{X}_\infty \leq -\zeta\}}| \right] = |\zeta|. \quad (2.15)$$

45 where, for a set F , $\mathbb{1}_F$ denotes the indicator function of F .

Proof of Lemma 2.3. The proof is similar to that in Appendix E.3.1. of Dai and Shi [5], and thus is omitted here. □

2.3. Bounding partial third moment (1.12)

Consider a function $V(x) = x^4 + a_1 x^3 + a_2 x^2$, where a_1 and a_2 are two constants that will be determined later. Recalling the definition of $G_{\tilde{X}}$ in the main paper, there is

$$\begin{aligned} G_{\tilde{X}} V(x) &= 4x^3 \delta \mathbb{E}_n(A_0 - D_0) + \{6\delta^2 \mathbb{E}_n[(A_0 - D_0)^2] + 3a_1 \delta \mathbb{E}_n(A_0 - D_0)\} x^2 \\ &\quad + \{4\delta^3 \mathbb{E}_n[(A_0 - D_0)^3] + 3a_1 \delta^2 \mathbb{E}_n[(A_0 - D_0)^2] + 2a_2 \delta \mathbb{E}_n(A_0 - D_0)\} x \\ &\quad + \delta^4 \mathbb{E}_n[(A_0 - D_0)^4] + a_1 \delta^3 \mathbb{E}_n[(A_0 - D_0)^3] + a_2 \delta^2 \mathbb{E}_n[(A_0 - D_0)^2], \end{aligned}$$

where $n \in \mathbb{N}$ is such that $x = \delta(n - x_\infty)$.

Now we determine a_1 and a_2 by taking them to satisfy

$$\begin{aligned} 6\delta^2 \mathbb{E}_N[(A_0 - D_0)^2] + 3a_1 \delta \mathbb{E}_N(A_0 - D_0) &= 0, \\ 4\delta^3 \mathbb{E}_N[(A_0 - D_0)^3] + 3a_1 \delta^2 \mathbb{E}_N[(A_0 - D_0)^2] + 2a_2 \delta \mathbb{E}_N(A_0 - D_0) &= 0. \end{aligned} \quad (2.16)$$

Then,

$$\begin{aligned}
a_1 &= 2\delta\mathbb{E}_N[(A_0 - D_0)^2]/\mathbb{E}_N(D_0 - A_0) \\
&= 2\delta \left[1 - \mu + (2 - \mu)\frac{\Lambda}{N\mu - \Lambda} + (N\mu - \Lambda) \right] \\
&= 2 \left[(1 - \mu)\delta + \mu|\zeta| + (2 - \mu)\frac{1}{|\zeta|} \right], \tag{2.17}
\end{aligned}$$

and

$$\begin{aligned}
a_2 &= \{4\delta^2\mathbb{E}_N[(A_0 - D_0)^3] + 3a_1\delta\mathbb{E}_N[(A_0 - D_0)^2]\} / 2\mathbb{E}_N(D_0 - A_0) \\
&\leq 2\delta^2 40 [1 \vee (N\mu)^2] \frac{\delta N}{|\zeta|} + \frac{3}{4}a_1^2 \\
&= 80 [1 \vee (N\mu)^2] \frac{N\delta^3}{|\zeta|} + \frac{3}{4}a_1^2. \tag{2.18}
\end{aligned}$$

50 where the second line uses the inequality (2.7).

With the a_1 and a_2 chosen as above,

$$\begin{aligned}
G_{\tilde{X}}V(x) &= 4x^3b(x) + \{6\delta^2\mathbb{E}_n[(A_0 - D_0)^2] + 3a_1\delta\mathbb{E}_n(A_0 - D_0)\} x^2\mathbb{1}_{\{x < -\zeta\}} \\
&\quad + \{4\delta^3\mathbb{E}_n[(A_0 - D_0)^3] + 3a_1\delta^2\mathbb{E}_n[(A_0 - D_0)^2] + 2a_2\delta\mathbb{E}_n(A_0 - D_0)\} x\mathbb{1}_{\{x < -\zeta\}} \\
&\quad + \delta^4\mathbb{E}_n[(A_0 - D_0)^4] + a_1\delta^3\mathbb{E}_n[(A_0 - D_0)^3] + a_2\delta^2\mathbb{E}_n[(A_0 - D_0)^2], \tag{2.19}
\end{aligned}$$

where

$$4x^3b(x) = -4\mu x^4\mathbb{1}_{\{x \leq -\zeta\}} - 4\mu|\zeta|x^3\mathbb{1}_{\{x > -\zeta\}}. \tag{2.20}$$

Now with a proof similar to that in Appendix E.4.2 of [5], we have the following basic adjoint relation (BAR)

$$\mathbb{E} \left[G_{\tilde{X}}V(\tilde{X}_\infty) \right] = 0. \tag{2.21}$$

Taking expectation with respect to \tilde{X}_∞ on both sides of (2.19), we obtain

$$\begin{aligned}
4\mu\mathbb{E}[\tilde{X}_\infty^4\mathbb{1}_{\{\tilde{X}_\infty \leq -\zeta\}}] &\leq 2N\mu \left\{ 3a_1\delta \left(\frac{4}{3} + \frac{8}{3}\delta^2 \right) + 2a_2\delta \left[\sqrt{\frac{4}{3} + \frac{8}{3}\delta^2} \wedge 2|\zeta| \right] \right\} \\
&\quad + 8[1 \vee (N\mu)] N\mu \left\{ 6\delta^2 \left(\frac{4}{3} + \frac{8}{3}\delta^2 \right) + 3a_1\delta^2 \left[\sqrt{\frac{4}{3} + \frac{8}{3}\delta^2} \wedge 2|\zeta| \right] + a_2\delta^2 \right\} \\
&\quad + 40[1 \vee (N\mu)^2] N\mu \left\{ 4\delta^3 \left[\sqrt{\frac{4}{3} + \frac{8}{3}\delta^2} \wedge 2|\zeta| \right] + a_1\delta^3 \right\} \\
&\quad + 280[1 \vee (N\mu)^3] N\mu\delta^4, \tag{2.22}
\end{aligned}$$

where we use the moment bounds (2.11), (2.12), (2.13) in this supplement and the inequalities in Remark 2 of Lemma 2.1.

Applying the assumption $R \geq 1$, or $\delta \leq 1$, (2.22) implies that

$$\begin{aligned} \mathbb{E}[\tilde{X}_\infty^4 \mathbf{1}_{\{\tilde{X}_\infty \leq -\zeta\}}] &\leq \{48[1 \vee (N\mu)] + 80[1 \vee (N\mu)^2] \delta + 70[1 \vee (N\mu)^3] \delta^2\} N\delta^2 \\ &\quad + a_1 \{6\delta^{-1} + 12[1 \vee (N\mu)](1 \wedge |\zeta|) + 10[1 \vee (N\mu)^2] \delta\} N\delta^2 \\ &\quad + a_2 \{2(1 \wedge |\zeta|)\delta^{-1} + 2[1 \vee (N\mu)]\} N\delta^2. \end{aligned} \quad (2.23)$$

Now, recall the characterization of μ and N in (1.10). Under the settings of the theorems, either $s \in [1/2, 1]$, $q \in [1/2, 1]$ or $s \geq 1$, $q \in [0, 1]$. In both cases, we have

$$|\zeta| = \beta R^{q-1/2} = \beta \delta^{1-2q}, \quad (2.24)$$

$$N\delta^2 = (N - R + R)/R = \beta R^{q-1} + 1 \leq \beta + 1, \quad (2.25)$$

$$N\mu = \gamma(N\delta^2)\delta^{2s-2} \leq \gamma(\beta + 1)\delta^{2s-2}. \quad (2.26)$$

$$\mu/|\zeta| = \frac{\gamma\delta^{2s}}{\beta\delta^{1-2q}} = \frac{\gamma}{\beta}\delta^{2s+2q-1} \leq \frac{\gamma}{\beta}\delta. \quad (2.27)$$

Substituting (2.24) - (2.26) into (2.17) and (2.18) implies that

$$a_1 \leq 2 \left(\delta + \gamma\delta^{2s}\beta\delta^{1-2q} + \frac{2}{\beta}\delta^{2q-1} \right) \leq 2(1 + \gamma\beta + 2/\beta) := C_5(\gamma, \beta), \quad (2.28)$$

$$\begin{aligned} a_2 &\leq 80[1 \vee \gamma^2(\beta + 1)^2]\delta^{4s-4} \frac{\delta}{|\zeta|}(\beta + 1) + \frac{3}{4}a_1^2 \\ &= 80\frac{1+\beta}{\beta}[1 \vee \gamma^2(\beta + 1)^2]\delta^{4s+2q-4} + \frac{3}{4}C_5(\gamma, \beta)^2 \\ &\leq C_6(\gamma, \beta)\delta^{\min(4s+2q-4, 0)}, \end{aligned} \quad (2.29)$$

where

$$C_6(\gamma, \beta) = 80\frac{1+\beta}{\beta}[1 \vee \gamma^2(\beta + 1)^2] + \frac{3}{4}C_5(\gamma, \beta)^2. \quad (2.30)$$

Next, substituting (2.24) - (2.26), and (2.28), (2.29) into (2.23), we obtain

$$\begin{aligned} \mathbb{E}[\tilde{X}_\infty^4 \mathbf{1}_{\{\tilde{X}_\infty \leq -\zeta\}}] &\leq (1 + \beta)\{48[1 \vee \gamma(1 + \beta)]\delta^{2s-2} + 80[1 \vee \gamma^2(1 + \beta)^2]\delta^{4s-3} + 70[1 \vee \gamma^3(1 + \beta)^3]\delta^{6s-4}\} \\ &\quad + C_5(\gamma, \beta)(1 + \beta)\{6 + 12[1 \vee \gamma(1 + \beta)] + 10[1 \vee \gamma^2(1 + \beta)^2]\}\delta^{-1} \\ &\quad + C_6(\gamma, \beta)(1 + \beta)\{2 + 2[1 \vee \gamma(1 + \beta)]\}\delta^{\min(4s+2q-4, 0)}\delta^{-1} \\ &\leq C_7(\gamma, \beta)\delta^{\min(4s+2q-4, 0)}\delta^{-1}, \end{aligned} \quad (2.31)$$

where

$$\begin{aligned}
C_7(\gamma, \beta) &= (1 + \beta) \{48[1 \vee \gamma(1 + \beta)] + 80[1 \vee \gamma^2(1 + \beta)^2] + 70[1 \vee \gamma^3(1 + \beta)^3]\} \\
&+ C_5(\gamma, \beta)(1 + \beta) \{6 + 12[1 \vee \gamma(1 + \beta)] + 10[1 \vee \gamma^2(1 + \beta)^2]\} \\
&+ C_6(\gamma, \beta)(1 + \beta) \{2 + 2[1 \vee \gamma(1 + \beta)]\}.
\end{aligned} \tag{2.32}$$

Applying Jensen's inequality to (2.31) proves Lemma 1.1 in Section 1 of this supplement.

3. Discrete-time queue with general arrival distribution

55 We consider the customer count process, $X = \{X_k : k = 0, 1, \dots\}$, with a general arrival distribution. Specifically, comparing with the discrete queue studied in the main paper [1], we assume that the arrivals $\{A_k : k = 0, 1, \dots\}$ form an i.i.d. sequence and follows a general distribution $G(\cdot)$ such that

- Variance of the distribution is $\sigma_A^2 < \infty$;
- 60 • Third non-central moment of the distribution is $\mu_3 < \infty$.

Note that the Poisson arrival case studied in the main paper is one special case satisfying the two conditions above.

Let X_∞ and \tilde{X}_∞ be the steady-state customer count and the scaled version of it, defined in Section 1 of the main paper. Let Y_∞ be the continuous random variable having the following density

$$p(x) \propto \frac{2}{a(x)} \exp \left(\int_0^x \frac{2b(y)}{a(y)} dy \right), \quad x \in \mathbb{R}, \tag{3.1}$$

where $b(x)$ is the same as the one defined by (1.5) of the main paper, and

$$a(x) = \begin{cases} \mu [(c_A - 1) + \Lambda], & x \leq -1/\delta, \\ \mu (c_A - \mu + \delta(1 - \mu)x + \mu x^2), & x \in [-1/\delta, |\zeta|], \\ \mu (c_A - \mu + \delta(1 - \mu)|\zeta| + \mu \zeta^2), & x \geq |\zeta|. \end{cases} \tag{3.2}$$

Here,

$$c_A \equiv \sigma_A^2 / \Lambda + 1. \tag{3.3}$$

Note that when the arrival distribution is Poisson, $c_A = 2$. In that case, (3.2) coincides with the definition of $a(x)$ in (1.6) of the main paper.

Note that $p(x)$ is the stationary density of the diffusion process

$$G_Y f(x) = b(x)f'(x) + \frac{1}{2}a(x)f''(x), \quad x \in \mathbb{R}, f \in C^2(\mathbb{R}). \quad (3.4)$$

65 Next, we state the main theorem for this discrete-time queueing system with a general arrival distribution.

Theorem 3.1. *Consider the DTMC X with arrivals $\{A_k : k = 0, 1, \dots\}$ following distribution $G(\cdot)$ such that*

$$\text{Var}(A_0)/\Lambda = c_A - 1 < \infty, \quad \mathbb{E}A_0^3/\Lambda = v_A < \infty. \quad (3.5)$$

For all $N \geq 1, \mu \in (0, 1)$ satisfying $1 \leq R < N$ and (1.10) for some $s \geq 1, \frac{1}{2} \leq q \leq 1$. The Wasserstein distance between \tilde{X}_∞ and Y_∞

$$d_W(\tilde{X}_\infty, Y_\infty) \leq C(\gamma, \beta, c_A, v_A)\delta, \quad (3.6)$$

where

$$C(\gamma, \beta, c_A, v_A) = C_0 \left(1 + \frac{1}{\beta}\right) \left[\frac{2}{3}v_A + \frac{10}{3}(1 + \beta) \left\{1 \vee (\gamma(1 + \beta))^2\right\}\right]. \quad (3.7)$$

Here, $C_0 = C_0(\gamma, c_A)$ is a constant depending only on γ and c_A , with the explicit form specified in Lemma 3.1. Note that in the Poisson arrival case, $c_A = 2$ and C_0 coincides with its counterpart in Lemma 2 of the main paper.

70 *Proof of Theorem 3.1.* The basic framework to prove Theorem 3.1 is the same as that in the main paper.

For any $h \in \text{Lip}(1)$, let $f = f_h$ be a solution to the Poisson equation

$$G_Y f(x) = \mathbb{E}[h(Y_\infty)] - h(x), \quad x \in \mathbb{R}. \quad (3.8)$$

After the generator coupling via the Poisson equation

$$\left| \mathbb{E}h(\tilde{X}_\infty) - \mathbb{E}h(Y_\infty) \right| = \left| \mathbb{E}[G_{\tilde{X}} f(x) - G_Y f(x)] \right|, \quad (3.9)$$

we perform the following Taylor expansion for any given $x = \delta(n - x_\infty)$ and $n = 0, 1, \dots$,

$$\begin{aligned} G_{\tilde{X}} f(x) &= \mathbb{E}_n[f(x + \delta(A_0 - D_0))] - f(x) \\ &= f'(x)\delta\mathbb{E}_n(A_0 - D_0) + \frac{1}{2}f''(x)\delta^2\mathbb{E}_n[(A_0 - D_0)^2] + \frac{1}{6}\delta^3\mathbb{E}_n[f'''(\xi)(A_0 - D_0)^3], \end{aligned} \quad (3.10)$$

where

$$|\xi - x| \leq \delta |A_0 - D_0|.$$

It can be easily verified that

$$\delta\mathbb{E}_n(A_0 - D_0) = b(x),$$

and

$$\delta^2\mathbb{E}_n[(A_0 - D_0)^2] = a(x), \quad (3.11)$$

where (3.11) follows from the calculation below

$$\begin{aligned} \delta^2\mathbb{E}_n[(A_0 - D_0)^2] &= \delta^2\text{Var}_n(A_0 - D_0) + \delta^2[\mathbb{E}_n(A_0 - D_0)]^2 \\ &= \delta^2\sigma_A^2 + \delta^2[N - (n - N)^-]\mu(1 - \mu) + b^2(x) \\ &= \delta^2\sigma_A^2 + [\delta^2N\mu - \delta(x + \zeta)^-\mu](1 - \mu) + b^2(x) \\ &= \delta^2\sigma_A^2 + [-\delta b(x) + \delta^2\Lambda](1 - \mu) + b^2(x) \\ &= \delta^2\sigma_A^2 + \delta^2\Lambda(1 - \mu) - (1 - \mu)\delta b(x) + b^2(x). \end{aligned} \quad (3.12)$$

Combining (3.9) and (3.10) implies that

$$\begin{aligned} \left| \mathbb{E}h(\tilde{X}_\infty) - \mathbb{E}h(Y_\infty) \right| &= \left| \mathbb{E}[G_{\tilde{X}}f(x) - G_Yf(x)] \right| \\ &\leq \frac{1}{6}\delta^3\mathbb{E}\left\{ \mathbb{E}_{X_\infty} \left[\|f'''\| |A_0 - D_0|^3 \right] \right\}, \end{aligned} \quad (3.13)$$

where $\|f'''\| = \max_{x \in \mathbb{R}} |f'''(x)|$.

For (3.13), one first uses the c_r -inequality and equation (2.3) to obtain

$$\mathbb{E}_n |A_0 - D_0|^3 \leq 4\mathbb{E}A_0^3 + 20 \max\{1, (N\mu)^2\} N\mu = 4\frac{\mu_3}{\Lambda} \Lambda + 20 \max\{1, (N\mu)^2\} N\mu, \quad (3.14)$$

for each $n = 0, 1, \dots$. Then, applying the gradient bound (3.20) stated in Lemma 3.1 below, we get

$$\begin{aligned} \frac{1}{6}\delta^3\mathbb{E}\left\{ \mathbb{E}_{X_\infty} \left[\|f'''\| |A_0 - D_0|^3 \right] \right\} &\leq \frac{1}{6}C_0(\mu\delta^{-1}, c_A)\delta^3\frac{1}{\mu}\left(1 + \frac{1}{|\zeta|}\right)\left[4\frac{\mu_3}{\Lambda}\Lambda + 20 \max\{1, (N\mu)^2\} N\mu\right] \\ &\leq \left[\frac{2}{3}\frac{\mu_3}{\Lambda} + \frac{10}{3} \max\{1, (N\mu)^2\}\frac{N}{R}\right]C_0(\mu\delta^{-1}, c_A)\left(1 + \frac{1}{|\zeta|}\right)\delta. \end{aligned} \quad (3.15)$$

Recall the characterizations of $|\zeta|$, $N\delta^2$, $N\mu$, and $\mu/|\zeta|$ in (2.24)-(2.27) of this supplement. Under the assumptions on s and q in Theorem 3.1, we have $|\zeta| \geq \beta$ and $N\mu \leq \gamma(\beta + 1)$. Applying these characterizations to (3.15), we obtain through (3.13) that

$$\begin{aligned} \left| \mathbb{E}h(\tilde{X}_\infty) - \mathbb{E}h(Y_\infty) \right| &= \left| \mathbb{E}G_{\tilde{X}}f(\tilde{X}_\infty) - \mathbb{E}G_Yf(\tilde{X}_\infty) \right| \\ &\leq C_0(\gamma, c_A) \left(1 + \frac{1}{\beta} \right) \left[\frac{2}{3}v_A + \frac{10}{3}(1 + \beta) \left\{ 1 \vee (\gamma(1 + \beta))^2 \right\} \right] \delta, \end{aligned} \quad (3.16)$$

where the inequality comes from the observation that $C_0(\cdot, c_A)$ is an increasing function in its first variable, and that $\mu\delta^{-1} \leq \gamma$.

75

This proves Theorem 3.1. \square

Lemma 3.1 (Gradient bounds). *Fix an $h \in \text{Lip}(1)$ with $h(0) = 0$. There exists a solution f_h to the Poisson equation,*

$$G_Y f(x) = \mathbb{E}h(Y) - h(x), \quad x \in \mathbb{R}, \quad (3.17)$$

that is twice continuously differentiable, with an absolutely continuous second derivative, and for all $\Lambda > 0$, $N \geq 1$, and $\mu \in (0, 1)$ satisfying $1 \leq R < N$,

$$|f'_h(x)| \leq \begin{cases} \frac{\tilde{C}_1}{\mu}(1 + 1/|\zeta|), & x \leq -\zeta, \\ \frac{1}{2} + \frac{1}{\mu|\zeta|} \left[x + \left(\tilde{C} + \frac{\delta}{2} \right) + \left(\tilde{C} + \frac{c_A}{2} \right) \frac{1}{|\zeta|} \right], & x \geq -\zeta, \end{cases} \quad (3.18)$$

$$|f''_h(x)| \leq \begin{cases} \frac{\tilde{C}_2}{\mu}(1 + 1/|\zeta|), & x \leq -\zeta, \\ \frac{1}{\mu|\zeta|}, & x \geq -\zeta, \end{cases} \quad (3.19)$$

$$|f'''_h(x)| \leq \begin{cases} \frac{C_0}{\mu}(1 + 1/|\zeta|), & x \leq -\zeta, \\ \frac{4}{c_A - 1} \frac{1}{\mu}, & x \geq -\zeta, \end{cases}, \quad (3.20)$$

where

$$\tilde{C}_1 = \tilde{C}_1(\mu\delta^{-1}, c_A) = \tilde{C}e^{\frac{1}{c_A-1}} \left(3 + \frac{2}{c_A - 1} + \frac{2}{c_A - 1} e^{\frac{1}{c_A-1}} \right), \quad (3.21)$$

$$\begin{aligned} \tilde{C}_2 = \tilde{C}_2(\mu\delta^{-1}, c_A) &= e^{\frac{1}{c_A-1}} \left(1 + \frac{2}{c_A - 1} + \frac{2}{c_A - 1} e^{\frac{1}{c_A-1}} \right) \left[1 + (1 + \tilde{C}) \left(\frac{\mu\delta^{-1}}{c_A - 1} \vee \frac{1}{c_A - 1} \vee 2 \right) \right. \\ &\quad \left. + \tilde{C}_1 \left(\frac{c_A - \mu}{c_A - 1} \vee 3 \right) \right], \end{aligned} \quad (3.22)$$

$$C_0 = C_0(\mu\delta^{-1}, c_A) = \frac{4}{c_A - 1} \left[1 + (1 + \tilde{C}) \left(\frac{\mu\delta^{-1}}{c_A - 1} \vee \frac{1}{c_A - 1} \vee 2 \right) + \tilde{C}_1 \left(\frac{c_A - \mu}{c_A - 1} \vee 3 \right) \right], \quad (3.23)$$

and

$$\begin{aligned}\tilde{C} &= \tilde{C}(\mu\delta^{-1}, c_A) \\ &= \frac{1+\sqrt{2}}{2}\mu\delta^{-1} (1 \vee \mu\delta^{-1}) + \left[2 + \sqrt{\mu\delta^{-1}} \left(1 \vee \sqrt{\mu\delta^{-1}}\right)\right] \left[c_A + \frac{3}{2}(1-\mu)\delta^2\right].\end{aligned}\quad (3.24)$$

We leave the complete details of the proof for these gradient bounds to the last section due to its complexities.

4. Gradient bounds for state-dependent diffusion process

To establish the gradient bounds in Lemma 3.1, we first define the following useful quantity for notational convenience

$$r(x) \equiv \frac{2b(x)}{a(x)} = \begin{cases} \frac{-2x}{(c_A-1)+\Lambda}, & x \leq -1/\delta, \\ \frac{-2x}{c_A-\mu+\delta(1-\mu)x+\mu x^2}, & x \in [-1/\delta, |\zeta|], \\ \frac{-2|\zeta|}{c_A-\mu+\delta(1-\mu)|\zeta|+\mu\zeta^2}, & x \geq |\zeta|. \end{cases}\quad (4.1)$$

Note that for $x \in [-1/\delta, 0]$,

$$r(x) \leq \frac{-2x}{c_A - \mu + \delta(1-\mu)(-1/\delta)} = \frac{-2x}{c_A - 1},\quad (4.2)$$

and for $x \in [0, -\zeta]$,

$$-r(x) \leq \frac{2x}{c_A - \mu} \leq \frac{2}{c_A - 1}x.\quad (4.3)$$

These inequalities will turn out to be of use in the proof later.

⁸⁰ Then, we need the following three lemmas as preliminaries, which will be proven at the end of this section.

Lemma 4.1. Recall that $q(x)$ is defined by (3.10) of the main paper [1]. The following bounds hold:

$$\frac{1}{q(x)} \int_{-\infty}^x \frac{2}{a(y)} q(y) dy \leq \begin{cases} \frac{1}{\mu}, & x \leq -1, \\ \left(1 + \frac{2}{c_A - 1} e^{\frac{1}{c_A - 1}}\right) \frac{1}{\mu}, & x \in [-1, 0], \\ e^{\frac{1}{c_A - 1} \zeta^2} \left(1 + \frac{2}{c_A - 1} e^{\frac{1}{c_A - 1}} + \frac{2}{c_A - 1} |\zeta|\right) \frac{1}{\mu}, & x \in [0, -\zeta]. \end{cases} \quad (4.4)$$

$$\frac{1}{q(x)} \int_x^{\infty} \frac{2}{a(y)} q(y) dy \leq \begin{cases} \left(\frac{1}{\eta} + \frac{2\eta}{c_A - 1} e^{\frac{\eta^2}{c_A - 1}}\right) \frac{1}{\mu}, & x \in [0, \eta], \eta \leq -\zeta, \\ \frac{1}{\mu|\zeta|}, & x \geq -\zeta. \end{cases} \quad (4.5)$$

$$\frac{1}{q(x)} \int_{-\infty}^x \frac{2|y|}{a(y)} q(y) dy \leq \begin{cases} \frac{1}{\mu}, & x \leq 0, \\ 2e^{\frac{1}{c_A - 1} \zeta^2} \frac{1}{\mu}, & x \in [0, -\zeta]. \end{cases} \quad (4.6)$$

$$\frac{1}{q(x)} \int_x^{\infty} \frac{2|y|}{a(y)} q(y) dy \leq \begin{cases} \frac{3}{2} \frac{1}{\mu} + \frac{\delta}{2} \frac{1}{\mu|\zeta|} + \frac{c_A}{2} \frac{1}{\mu\zeta^2}, & x \in [0, -\zeta], \\ \frac{x}{\mu|\zeta|} + \frac{\delta}{2} \frac{1}{\mu|\zeta|} + \frac{c_A}{2} \frac{1}{\mu\zeta^2} + \frac{1}{2}, & x \geq -\zeta. \end{cases} \quad (4.7)$$

$$\frac{|r(x)|}{q(x)} \int_{-\infty}^x \frac{2}{a(y)} q(y) dy \leq \frac{2}{c_A - 1} \frac{1}{\mu}, \quad x \leq 0. \quad (4.8)$$

$$\frac{|r(x)|}{q(x)} \int_x^{\infty} \frac{2}{a(y)} q(y) dy \leq \frac{2}{c_A - 1} \frac{1}{\mu}, \quad x \geq 0, \quad (4.9)$$

and when $|\zeta| \geq 1$,

$$\frac{1}{q(x)} \int_x^{\infty} \frac{2}{a(y)} q(y) dy \leq \begin{cases} \left(1 + \frac{2}{c_A - 1} e^{\frac{1}{c_A - 1}}\right) \frac{1}{\mu}, & x \in [0, 1], \\ \frac{1}{\mu}, & x \geq 1. \end{cases} \quad (4.10)$$

Lemma 4.2. Let the random variable Y_{∞} have the stationary distribution of a diffusion process with drift $b(x)$ and state-dependent diffusion coefficient $a(x)$.

$$\mathbb{E}|Y_{\infty}| \leq \tilde{C} (1 + 1/|\zeta|), \quad (4.11)$$

where \tilde{C} is specified by (3.24) in Lemma 3.1.

Lemma 4.3. Recall the form of $a(x)$ and $r(x)$ in (3.2) and (4.1). The following bounds hold:

$$a(x) \geq (c_A - 1)\mu, \quad x \in \mathbb{R}; \quad (4.12)$$

$$\frac{|xa'(x)|}{a(x)} \leq \left(\frac{1-\mu}{c_A-1} \vee 2 \right) \mathbf{1}_{\{x \in (-1/\delta, -\zeta]\}}, \quad x \in \mathbb{R}; \quad (4.13)$$

$$\mathbb{E} |Y_\infty| \frac{|a'(x)|}{a(x)} \leq \tilde{C} \left(\frac{\mu\delta^{-1}}{c_A-1} \vee \frac{1}{c_A-1} \vee 2 \right) (1 + 1/|\zeta|) \mathbf{1}_{\{x \in (-1/\delta, -\zeta]\}}, \quad x \in \mathbb{R}; \quad (4.14)$$

$$|r'(x)a(x)| \leq 2 \left(\frac{c_A - \mu}{c_A - 1} \vee 3 \right) \mu \mathbf{1}_{\{x \leq -\zeta\}}, \quad x \in \mathbb{R}; \quad (4.15)$$

where $a'(x)$ and $r'(x)$ are interpreted as the left derivative at $x = 1/\delta$ and $x = -\zeta$.

Now we are going to prove Lemma 3.1 using these three lemmas.

Following from the paper by Braverman and Dai [6], the derivatives of $f_h(x)$ have the following forms:

$$f'_h(x) = \frac{1}{q(x)} \int_{-\infty}^x \frac{2}{a(y)} (\mathbb{E}h(Y_\infty) - h(y)) q(y) dy, \quad x \in \mathbb{R}. \quad (4.16)$$

$$f'_h(x) = -\frac{1}{q(x)} \int_x^{\infty} \frac{2}{a(y)} (\mathbb{E}h(Y_\infty) - h(y)) q(y) dy, \quad x \in \mathbb{R}. \quad (4.17)$$

$$f''_h(x) = \frac{1}{q(x)} \int_{-\infty}^x \left\{ -\frac{2}{a(y)} h'(y) - \frac{2a'(y)}{a^2(y)} [\mathbb{E}h(Y_\infty) - h(y)] - r'(y) f'_h(y) \right\} q(y) dy, \quad (4.18)$$

$$f''_h(x) = \frac{1}{q(x)} \int_x^{\infty} \left\{ -\frac{2}{a(y)} h'(y) - \frac{2a'(y)}{a^2(y)} [\mathbb{E}h(Y_\infty) - h(y)] - r'(y) f'_h(y) \right\} q(y) dy, \quad (4.19)$$

$$f'''_h(x) = -r'(x) f'_h(x) - r(x) f''_h(x) - \frac{2}{a(x)} h'(x) - \frac{2a'(x)}{a^2(x)} [\mathbb{E}h(Y_\infty) - h(x)], \quad (4.20)$$

85 where $a'(x)$ is interpreted as the left derivative at the points $x = -1/\delta$ and $x = -\zeta$.

Then, the properties of h implies that

$$|f'_h(x)| \leq \frac{1}{q(x)} \int_{-\infty}^x \frac{2}{a(y)} (\mathbb{E} |Y_\infty| + |y|) q(y) dy, \quad (4.21)$$

$$|f'_h(x)| \leq \frac{1}{q(x)} \int_x^{\infty} \frac{2}{a(y)} (\mathbb{E} |Y_\infty| + |y|) q(y) dy. \quad (4.22)$$

For $x \leq 0$, applying (4.4), (4.6), and (4.11) to (4.21) gives

$$|f'_h(x)| \leq \frac{1}{\mu} (1 + 1/|\zeta|) \left[1 + \tilde{C} \left(1 + \frac{2}{c_A - 1} e^{\frac{1}{c_A - 1}} \right) \right]. \quad (4.23)$$

For $x \in [0, -\zeta]$, we need to consider separately the cases when $|\zeta| \leq 1$ and $|\zeta| \geq 1$.

When $|\zeta| \leq 1$, applying (4.4), (4.6) and (4.11) to (4.21) gives

$$|f'_h(x)| \leq \frac{1}{\mu} (1 + 1/|\zeta|) \tilde{C} \left[e^{\frac{1}{c_A - 1}} \left(3 + \frac{2}{c_A - 1} + \frac{2}{c_A - 1} e^{\frac{1}{c_A - 1}} \right) \right], \quad x \in [0, -\zeta], \quad (4.24)$$

and when $|\zeta| \geq 1$, applying (4.5), (4.10) and (4.11) to (4.22) gives

$$|f'_h(x)| \leq \frac{1}{\mu}(1 + 1/|\zeta|) \left[\tilde{C} \left(1 + \frac{2}{c_A - 1} e^{\frac{1}{c_A - 1}} \right) + 1 + \frac{c_A}{2} \right], \quad x \in [0, -\zeta]. \quad (4.25)$$

Using (4.23)-(4.25) together, along with the observation that $2\tilde{C} > 1 + \frac{c_A}{2}$, proves the first half of (3.18).

For $x \geq -\zeta$, applying (4.5), (4.7) and (4.11) to (4.22) gives

$$|f'_h(x)| \leq \frac{1}{2} + \frac{1}{\mu|\zeta|} \left[x + \left(\tilde{C} + \frac{\delta}{2} \right) + \left(\tilde{C} + \frac{c_A}{2} \right) \frac{1}{|\zeta|} \right], \quad (4.26)$$

which proves the second half of (3.18).

90 Now, we move on to deal with (3.19) and (3.20).

Since the function h satisfies $|h(x)| \leq |x|$ for all $x \in \mathbb{R}$ and $\|h'\| \leq 1$, (4.18) and (4.19) imply that

$$|f''_h(x)| \leq \frac{1}{q(x)} \int_{-\infty}^x \left\{ \frac{2}{a(y)} + \mathbb{E}|Y_\infty| \frac{2|a'(y)|}{a^2(y)} + \frac{2|ya'(y)|}{a^2(y)} + |r'(y)f'_h(y)| \right\} q(y) dy, \quad (4.27)$$

$$|f''_h(x)| \leq \frac{1}{q(x)} \int_x^\infty \left\{ \frac{2}{a(y)} + \mathbb{E}|Y_\infty| \frac{2|a'(y)|}{a^2(y)} + \frac{2|ya'(y)|}{a^2(y)} + |r'(y)f'_h(y)| \right\} q(y) dy. \quad (4.28)$$

Thus, when $x \leq -\zeta$, applying (4.14), (4.13), (4.15) and (3.18) to (4.27) implies that

$$\begin{aligned} |f''_h(x)| &\leq \frac{1}{q(x)} \int_{-\infty}^x \frac{2}{a(y)} q(y) \left\{ 1 + \tilde{C} \left(\frac{\mu\delta^{-1}}{c_A - 1} \vee \frac{1}{c_A - 1} \vee 2 \right) (1 + 1/|\zeta|) \mathbb{1}_{\{y \in (-1/\delta, -\zeta]\}} \right. \\ &\quad \left. + \left(\frac{1 - \mu}{c_A - 1} \vee 2 \right) \mathbb{1}_{\{y \in (-1/\delta, -\zeta]\}} + \tilde{C}_1 \left(\frac{c_A - \mu}{c_A - 1} \vee 3 \right) (1 + 1/|\zeta|) \mathbb{1}_{\{y \leq -\zeta\}} \right\} dy \\ &\leq \hat{C}(1 + 1/|\zeta|) \frac{1}{q(x)} \int_{-\infty}^x \frac{2}{a(y)} q(y) dy, \end{aligned} \quad (4.29)$$

where

$$\hat{C} = 1 + (1 + \tilde{C}) \left(\frac{\mu\delta^{-1}}{c_A - 1} \vee \frac{1}{c_A - 1} \vee 2 \right) + \tilde{C}_1 \left(\frac{c_A - \mu}{c_A - 1} \vee 3 \right);$$

and when $x \geq 0$, we again apply (4.14), (4.13), (4.15) and (3.18), but this time to (4.28), to see that

$$\begin{aligned} |f''_h(x)| &\leq \frac{1}{q(x)} \int_x^\infty \frac{2}{a(y)} q(y) \left\{ 1 + \tilde{C} \left(\frac{\mu\delta^{-1}}{c_A - 1} \vee \frac{1}{c_A - 1} \vee 2 \right) (1 + 1/|\zeta|) \mathbb{1}_{\{y \in (-1/\delta, -\zeta]\}} \right. \\ &\quad \left. + \left(\frac{1 - \mu}{c_A - 1} \vee 2 \right) \mathbb{1}_{\{y \in (-1/\delta, -\zeta]\}} + \tilde{C}_1 \left(\frac{c_A - \mu}{c_A - 1} \vee 3 \right) (1 + 1/|\zeta|) \mathbb{1}_{\{y \leq -\zeta\}} \right\} dy \\ &\leq \left[1 + (\hat{C} - 1)(1 + 1/|\zeta|) \mathbb{1}_{\{x < -\zeta\}} \right] \frac{1}{q(x)} \int_x^\infty \frac{2}{a(y)} q(y) dy, \end{aligned} \quad (4.30)$$

where for the last inequality we used that for all $y > x$, $\mathbb{1}_{\{y \leq -\zeta\}} \leq \mathbb{1}_{\{x < -\zeta\}}$.

Therefore, when $x \leq 0$, applying (4.4) to (4.29) implies that

$$|f_h''(x)| \leq \frac{\hat{C}}{\mu}(1 + 1/|\zeta|) \left(1 + \frac{2}{c_A - 1} e^{\frac{1}{c_A - 1}}\right). \quad (4.31)$$

For $x \in [0, -\zeta]$, we need to consider separately the cases when $|\zeta| \leq 1$ and $|\zeta| \geq 1$. When $|\zeta| \leq 1$, applying (4.4) to (4.29) implies that

$$|f_h''(x)| \leq \frac{\hat{C}}{\mu}(1 + 1/|\zeta|) \left[e^{\frac{1}{c_A - 1}} \left(1 + \frac{2}{c_A - 1} + \frac{2}{c_A - 1} e^{\frac{1}{c_A - 1}}\right) \right], \quad x \in [0, -\zeta], \quad (4.32)$$

and when $|\zeta| \geq 1$, applying (4.10) to (4.30) implies that

$$|f_h''(x)| \leq \frac{\hat{C}}{\mu}(1 + 1/|\zeta|) \left(1 + \frac{2}{c_A - 1} e^{\frac{1}{c_A - 1}}\right), \quad x \in [0, -\zeta]. \quad (4.33)$$

Combining the bounds in (4.31)-(4.33) proves the first half of (3.19).

When $x \geq -\zeta$, applying (4.5) to (4.30) immediately establishes the remaining half of (3.19).

Now we move on to (3.20). From the form of $f_h'''(x)$ in (4.20), along with the properties of the function h , we see immediately that

$$|f_h'''(x)| \leq |r'(x)f_h'(x)| + |r(x)f_h''(x)| + \frac{2}{a(x)} + \frac{2|a'(x)|}{a^2(x)} \mathbb{E}|Y_\infty| + \frac{2|xa'(x)|}{a^2(x)}.$$

Applying the bound on $|r'(x)|$ in (4.15), the bound on $|f_h'(x)|$ in (3.18), and the bounds (4.12)-(4.14), we have that

$$|f_h'''(x)| \leq \frac{2}{c_A - 1} \frac{1}{\mu} \left[1 + (\hat{C} - 1)(1 + 1/|\zeta|) \mathbb{1}_{\{x \leq -\zeta\}} \right] + |r(x)f_h''(x)|. \quad (4.34)$$

For the last term of the right side above, $|r(x)f_h''(x)|$, one can simply multiply both sides of (4.29) and (4.30) by $|r(x)|$, and then apply (4.8) and (4.9) to arrive at

$$|r(x)f_h''(x)| \leq \begin{cases} \frac{2}{c_A - 1} \frac{\hat{C}}{\mu} (1 + 1/|\zeta|), & x \leq -\zeta, \\ \frac{2}{c_A - 1} \frac{1}{\mu}, & x \geq -\zeta. \end{cases} \quad (4.35)$$

Combining (4.34) and (4.35) proves (3.20).

95 Finally, we are going to verify the three lemmas stated at the beginning of this section.

Proof of Lemma 4.1. First, we claim that

$$\frac{1}{q(x)} \int_{-\infty}^x \frac{2}{a(y)} q(y) dy \leq \frac{1}{b(x)}, \quad x < 0, \quad (4.36)$$

$$\frac{1}{q(x)} \int_x^{\infty} \frac{2}{a(y)} q(y) dy \leq \frac{1}{|b(x)|}, \quad x > 0. \quad (4.37)$$

To see why, suppose that $x < 0$. Using the fact that $b(y)/b(x) \geq 1$ for $y \leq x$, there is

$$\begin{aligned}
\frac{1}{q(x)} \int_{-\infty}^x \frac{2}{a(y)} q(y) dy &\leq \frac{1}{q(x)} \int_{-\infty}^x \frac{2b(y)}{a(y)} \frac{1}{b(x)} q(y) dy \\
&= \frac{1}{q(x)} \frac{1}{b(x)} \int_{-\infty}^x r(y) e^{\int_0^y r(u) du} dy \\
&= \frac{1}{q(x)} \frac{1}{b(x)} (q(x) - q(-\infty)) \\
&\leq \frac{1}{b(x)}.
\end{aligned} \tag{4.38}$$

The proof for (4.37) is essentially the same and is omitted here.

These two inequalities imply immediately the first part of (4.4) and the second part of (4.5).

It remains to bound the integrals when $x \in [-1, 0]$, and $x \in [0, -\zeta]$.

When $x \in [-1, 0]$,

$$\begin{aligned}
\frac{1}{q(x)} \int_{-\infty}^x \frac{2}{a(y)} q(y) dy &= \frac{q(-1)}{q(x)} \frac{1}{q(-1)} \int_{-\infty}^{-1} \frac{2}{a(y)} q(y) dy + \frac{1}{q(x)} \int_{-1}^x \frac{2}{a(y)} q(y) dy \\
&\leq \frac{q(-1)}{q(x)} \frac{1}{\mu} + \frac{1}{q(x)} \int_{-1}^x \frac{2}{a(y)} q(y) dy.
\end{aligned}$$

Observe that

$$\frac{q(-1)}{q(x)} = e^{\int_0^{-1} r(u) du - \int_0^x r(u) du} = e^{-\int_{-1}^x r(u) du} \leq 1.$$

Furthermore, using the inequality about $r(x)$ from (4.2) and the inequality $a(x) \geq (c_A - 1)\mu$ for all $x \in \mathbb{R}$ from (4.12), we see that

$$\begin{aligned}
\frac{1}{q(x)} \int_{-1}^x \frac{2}{a(y)} q(y) dy &= e^{\int_x^0 r(u) du} \int_{-1}^x \frac{2}{a(y)} e^{-\int_y^0 r(u) du} dy \\
&\leq e^{\int_x^0 r(u) du} \int_{-1}^0 \frac{2}{a(y)} dy \\
&\leq e^{\int_x^0 \frac{2}{c_A - 1} (-u) du} \frac{2}{(c_A - 1)\mu} \\
&\leq e^{\frac{1}{c_A - 1}} \frac{2}{(c_A - 1)\mu}.
\end{aligned}$$

Hence, for $x \in [-1, 0]$,

$$\frac{1}{q(x)} \int_{-\infty}^x \frac{2}{a(y)} q(y) dy \leq \frac{1}{\mu} \left(1 + \frac{2}{c_A - 1} e^{\frac{1}{c_A - 1}} \right).$$

This proves the second part of (4.4).

Now fix $\eta > 0$ such that $\eta \leq |\zeta|$. When $x \in [0, \eta]$,

$$\begin{aligned} \frac{1}{q(x)} \int_x^\infty \frac{2}{a(y)} q(y) dy &= \frac{1}{q(x)} \int_x^\eta \frac{2}{a(y)} q(y) dy + \frac{p(\eta)}{q(x)} \frac{1}{p(\eta)} \int_\eta^\infty \frac{2}{a(y)} q(y) dy \\ &\leq \frac{1}{q(x)} \int_x^\eta \frac{2}{a(y)} q(y) dy + \frac{p(\eta)}{q(x)} \frac{1}{\mu |\eta|}. \end{aligned}$$

To bound the first term above, observe that $r(x) < 0$ and $|r(x)| \leq \frac{2}{c_A - 1} x$ for $x \in [0, -\zeta]$. Then

$$\begin{aligned} \frac{1}{q(x)} \int_x^\eta \frac{2}{a(y)} q(y) dy &= e^{-\int_0^x r(u) du} \int_x^\eta \frac{2}{a(y)} e^{\int_0^y r(u) du} dy \\ &\leq e^{\eta^2/(c_A - 1)} \int_x^\eta \frac{2}{a(y)} dy \\ &\leq e^{\eta^2/(c_A - 1)} \frac{2\eta}{(c_A - 1)\mu}. \end{aligned}$$

Furthermore,

$$\frac{q(\eta)}{q(x)} = e^{\int_x^\eta r(u) du} \leq 1.$$

Hence when $x \in [0, \eta]$,

$$\frac{1}{q(x)} \int_x^\infty \frac{2}{a(y)} q(y) dy \leq e^{\eta^2/(c_A - 1)} \frac{2\eta}{c_A - 1} \frac{1}{\mu} + \frac{1}{\mu |\eta|}. \quad (4.39)$$

100 This proves the first part of (4.5).

Lastly, we are going to bound $\frac{1}{q(x)} \int_{-\infty}^x \frac{2}{a(y)} q(y) dy$ for $x \in [0, -\zeta]$. As before, observe that

$$\begin{aligned} \frac{1}{q(x)} \int_{-\infty}^x \frac{2}{a(y)} q(y) dy &= \frac{q(0)}{q(x)} \frac{1}{q(0)} \int_{-\infty}^0 \frac{2}{a(y)} q(y) dy + \frac{1}{q(x)} \int_0^x \frac{2}{a(y)} q(y) dy \\ &\leq \frac{q(0)}{q(x)} \frac{1}{\mu} \left(1 + \frac{2}{c_A - 1} e^{\frac{1}{c_A - 1}} \right) + \frac{2}{(c_A - 1)\mu} \frac{1}{q(x)} \int_0^x q(y) dy \\ &\leq e^{\frac{1}{c_A - 1} \zeta^2} \frac{1}{\mu} \left(1 + \frac{2}{c_A - 1} e^{\frac{1}{c_A - 1}} \right) + \frac{2}{(c_A - 1)\mu} e^{\frac{1}{c_A - 1} \zeta^2} |\zeta|, \end{aligned}$$

where the last inequality comes from

$$\frac{q(0)}{q(x)} = \exp \left(\int_0^x -r(y) dy \right) \leq \exp \left(\int_0^x \frac{2}{c_A - 1} y dy \right) = e^{\frac{1}{c_A - 1} x^2}.$$

This proves the last part of (4.4).

Therefore, (4.4) and (4.5) hold true.

Note that when $|\zeta| \geq 1$, taking $\eta = 1$ in (4.39) gives the first part of (4.10), while (4.37) gives the second part of it. This proves (4.10).

We move on to (4.6). For $x \leq 0$,

$$\frac{1}{q(x)} \int_{-\infty}^x \frac{2|y|}{a(y)} q(y) dy = \frac{1}{\mu} \frac{1}{q(x)} \int_{-\infty}^x r(y) q(y) dy \leq \frac{1}{\mu},$$

where the last inequality comes from

$$\int_{-\infty}^x r(y) q(y) dy = \int_{-\infty}^x r(y) e^{\int_0^y r(u) du} dy = q(x) - q(-\infty).$$

When $x \in [0, -\zeta]$,

$$\begin{aligned} \frac{1}{q(x)} \int_{-\infty}^x \frac{2|y|}{a(y)} q(y) dy &= \frac{1}{\mu} \frac{q(0)}{q(x)} \frac{1}{q(0)} \int_{-\infty}^0 r(y) q(y) dy - \frac{1}{\mu} \frac{1}{q(x)} \int_0^x r(y) q(y) dy \\ &= \frac{1}{\mu} \frac{q(0)}{q(x)} \frac{1}{q(0)} (q(0) - q(-\infty)) - \frac{1}{\mu} \frac{1}{q(x)} (q(x) - q(0)) \\ &\leq \frac{2}{\mu} \frac{q(0)}{q(x)} = \frac{2}{\mu} e^{\int_0^x |r(y)| dy} \leq \frac{2}{\mu} e^{\frac{1}{c_A - 1} \zeta^2}. \end{aligned}$$

105 This concludes the proof of (4.6).

We proceed to (4.7). For $x \in [0, -\zeta]$,

$$\begin{aligned} \frac{1}{q(x)} \int_x^\infty \frac{2|y|}{a(y)} q(y) dy &= -\frac{1}{\mu} \frac{1}{q(x)} \int_x^{|\zeta|} r(y) q(y) dy - \frac{1}{\mu} \frac{1}{|\zeta|} \frac{1}{q(x)} \int_{|\zeta|}^\infty y r(y) q(y) dy \\ &= \frac{1}{\mu} \left(1 - \frac{q(|\zeta|)}{q(x)} \right) - \frac{1}{\mu} \frac{1}{|\zeta|} \frac{1}{q(x)} \int_{|\zeta|}^\infty y r(y) q(y) dy \\ &= \frac{1}{\mu} \left(1 - \frac{q(|\zeta|)}{q(x)} \right) - \frac{1}{\mu} \frac{1}{|\zeta|} \frac{1}{q(x)} \left[-|\zeta| q(|\zeta|) - \int_{|\zeta|}^\infty q(y) dy \right] \\ &= \frac{1}{\mu} + \frac{1}{\mu} \frac{1}{|\zeta|} \frac{1}{q(x)} \int_{|\zeta|}^\infty q(y) dy \\ &\leq \frac{1}{\mu} + \frac{1}{\mu} \frac{1}{|\zeta|} \frac{c_A - \mu + \delta(1 - \mu) |\zeta| + \mu \zeta^2}{2 |\zeta|} \frac{q(|\zeta|)}{q(x)} \\ &\leq \frac{3}{2} \frac{1}{\mu} + \frac{1}{\mu} \frac{1}{|\zeta|} \frac{c_A + \delta |\zeta|}{2 |\zeta|}, \end{aligned}$$

where in the last inequality we use the fact that for $x \geq -\zeta$,

$$\begin{aligned} \frac{1}{q(x)} \int_x^\infty q(y) dy &= \int_x^\infty e^{\int_x^y r(u) du} dy = \int_0^\infty e^{\frac{-2|\zeta|}{c_A - \mu + \delta(1 - \mu) |\zeta| + \mu \zeta^2} y} dy \\ &= \frac{c_A - \mu + \delta(1 - \mu) |\zeta| + \mu \zeta^2}{2 |\zeta|}, \end{aligned} \tag{4.40}$$

and

$$\frac{q(|\zeta|)}{q(x)} = e^{\int_x^{|\zeta|} r(u) du} \leq 1.$$

For $x \geq -\zeta$,

$$\begin{aligned}
\frac{1}{q(x)} \int_x^\infty \frac{2|y|}{a(y)} q(y) dy &= -\frac{1}{\mu|\zeta|} \frac{1}{q(x)} \int_x^\infty y r(y) q(y) dy \\
&= -\frac{1}{\mu|\zeta|} \frac{1}{q(x)} \left[-xq(x) - \int_x^\infty q(y) dy \right] \\
&= \frac{x}{\mu|\zeta|} + \frac{1}{\mu|\zeta|} \frac{1}{q(x)} \int_x^\infty q(y) dy \\
&= \frac{x}{\mu|\zeta|} + \frac{1}{\mu|\zeta|} \frac{c_A - \mu + \delta(1-\mu)|\zeta| + \mu\zeta^2}{2|\zeta|}.
\end{aligned}$$

This proves (4.7).

Finally, we deal with (4.8) and (4.9). For $x < 0$ we use (4.36) to see that

$$\frac{|r(x)|}{q(x)} \int_{-\infty}^x \frac{2}{a(y)} q(y) dy \leq \frac{|r(x)|}{b(x)} = \frac{2}{a(x)} \leq \frac{2}{(c_A - 1)\mu}.$$

Similarly, we invoke (4.37) to see that when $x \geq 0$,

$$\frac{|r(x)|}{q(x)} \int_x^\infty \frac{2}{a(y)} q(y) dy \leq \frac{|r(x)|}{|b(x)|} = \frac{2}{a(x)} \leq \frac{2}{(c_A - 1)\mu}.$$

This proves (4.8) and (4.9), concluding our proof of Lemma 4.1. \square

Proof of Lemma 4.2. Consider the Lyapunov function $V(x) = x^2$. Using the form of G_Y in (3.4), we see immediately that

$$G_Y V(x) = 2xb(x) + a(x). \quad (4.41)$$

Recall the form of $b(x)$ and $a(x)$. When $x \leq -1/\delta$,

$$G_Y V(x) = 2x(-\mu x) + \mu[(c_A - 1) + \Lambda] = -2\mu x^2 + \mu(c_A - 1) + \mu^2 \delta^{-2},$$

where for the last equality we use $\delta = 1/\sqrt{R}$.

When $x \in [-1/\delta, -\zeta]$,

$$\begin{aligned}
G_Y V(x) &= -2\mu(1 - \mu/2)x^2 + \delta(1 - \mu)\mu x + \mu(c_A - \mu) \\
&\leq -2\mu(1 - \mu/2)x^2 + \mu \frac{1 - \mu}{2} x^2 + \mu \frac{1 - \mu}{2} \delta^2 + \mu(c_A - \mu) \\
&= -\frac{1}{2}(3 - \mu)\mu x^2 + \mu \frac{1 - \mu}{2} \delta^2 + \mu(c_A - \mu) \\
&\leq -\mu x^2 + \mu \frac{1 - \mu}{2} \delta^2 + \mu(c_A - \mu),
\end{aligned}$$

where to get the last inequality we use $\mu < 1$.

When $x \geq -\zeta$,

$$G_Y V(x) = -2\mu |\zeta| x + \mu(c_A - \mu) + \mu [\delta(1 - \mu) |\zeta| + \mu \zeta^2].$$

Therefore,

$$\begin{aligned} G_Y V(x) &\leq -2\mu x^2 \mathbf{1}_{\{x < -1/\delta\}} - \mu x^2 \mathbf{1}_{\{x \in [-1/\delta, -\zeta)\}} - 2\mu |\zeta| x \mathbf{1}_{\{x \geq -\zeta\}} \\ &\quad + \mu c_A + \mu^2 \delta^{-2} \mathbf{1}_{\{x < -1/\delta\}} + \frac{1 - \mu}{2} \delta^2 \mu \mathbf{1}_{\{x \in [-1/\delta, -\zeta)\}} \\ &\quad + \delta(1 - \mu) \mu |\zeta| \mathbf{1}_{\{x \geq -\zeta\}} + \mu^2 \zeta^2 \mathbf{1}_{\{x \geq -\zeta\}}. \end{aligned}$$

According to the standard Foster-Lyapunov criterion [7], for any $U, g_1, g_2 : \mathbb{R} \rightarrow \mathbb{R}_+$ satisfying

$$G_Y U(x) \leq -g_1(x) + g_2(x), \quad x \in \mathbb{R},$$

there is

$$\mathbb{E} g_1(Y_\infty) \leq \mathbb{E} g_2(Y_\infty).$$

Thus,

$$\begin{aligned} &2\mathbb{E} [Y_\infty^2 \mathbf{1}_{\{Y_\infty < -1/\delta\}}] + \mathbb{E} [Y_\infty^2 \mathbf{1}_{\{Y_\infty \in [-1/\delta, -\zeta)\}}] + 2|\zeta| \mathbb{E} [Y_\infty \mathbf{1}_{\{Y_\infty \geq -\zeta\}}] \\ &\leq c_A + \frac{1 - \mu}{2} \delta^2 + \mu \delta^{-2} \mathbb{P}(Y_\infty < -1/\delta) + (1 - \mu) \delta |\zeta| \mathbb{P}(Y_\infty \geq -\zeta) + \mu \zeta^2 \mathbb{P}(Y_\infty \geq -\zeta). \end{aligned} \quad (4.42)$$

Note that

$$\delta^{-1} \mathbb{P}(Y_\infty < -1/\delta) \leq \mathbb{E} [|Y_\infty| \mathbf{1}_{\{Y_\infty < -1/\delta\}}], \quad |\zeta| \mathbb{P}(Y_\infty \geq -\zeta) \leq \mathbb{E} [Y_\infty \mathbf{1}_{\{Y_\infty \geq -\zeta\}}].$$

Applying the above inequalities to (4.42), we get that

$$\begin{aligned} &2\mathbb{E} [Y_\infty^2 \mathbf{1}_{\{Y_\infty < -1/\delta\}}] + \mathbb{E} [Y_\infty^2 \mathbf{1}_{\{Y_\infty \in [-1/\delta, -\zeta)\}}] + 2|\zeta| \mathbb{E} [Y_\infty \mathbf{1}_{\{Y_\infty \geq -\zeta\}}] \\ &\leq c_A + \frac{1 - \mu}{2} \delta^2 + \mu \delta^{-1} \mathbb{E} [|Y_\infty| \mathbf{1}_{\{Y_\infty < -1/\delta\}}] + \mu |\zeta| \mathbb{E} [Y_\infty \mathbf{1}_{\{Y_\infty \geq -\zeta\}}] \\ &\quad + (1 - \mu) \delta^2 \mathbf{1}_{\{|\zeta| \leq \delta\}} + (1 - \mu) \delta \mathbb{E} [|Y_\infty| \mathbf{1}_{\{Y_\infty \geq -\zeta\}}] \mathbf{1}_{\{|\zeta| > \delta\}}. \end{aligned} \quad (4.43)$$

Since

$$2|\zeta| - \mu |\zeta| - (1 - \mu) \delta \mathbf{1}_{\{|\zeta| > \delta\}} \geq (2 - \mu) |\zeta| - (1 - \mu) |\zeta| = |\zeta|,$$

(4.43) implies that

$$\begin{aligned} & 2\mathbb{E} [Y_\infty^2 \mathbf{1}_{\{Y_\infty < -1/\delta\}}] + \mathbb{E} [Y_\infty^2 \mathbf{1}_{\{Y_\infty \in [-1/\delta, -\zeta]\}}] + |\zeta| \mathbb{E} [Y_\infty \mathbf{1}_{\{Y_\infty \geq -\zeta\}}] \\ & \leq c_A + \frac{3}{2}(1 - \mu)\delta^2 + \mu\delta^{-1}\mathbb{E} [|Y_\infty| \mathbf{1}_{\{Y_\infty < -1/\delta\}}]. \end{aligned} \quad (4.44)$$

From Jensen's inequality and (4.44), we have

$$\begin{aligned} \mathbb{E} [|Y_\infty| \mathbf{1}_{\{Y_\infty < -1/\delta\}}] & \leq \sqrt{\mathbb{E} [Y_\infty^2 \mathbf{1}_{\{Y_\infty < -1/\delta\}}]} \\ & \leq \sqrt{\frac{1}{2} \left[c_A + \frac{3}{2}(1 - \mu)\delta^2 + \mu\delta^{-1}\mathbb{E} [|Y_\infty| \mathbf{1}_{\{Y_\infty < -1/\delta\}}] \right]}, \end{aligned}$$

which is equivalent to a quadratic inequality in $\mathbb{E} [|Y_\infty| \mathbf{1}_{\{Y_\infty < -1/\delta\}}]$,

$$2 \left\{ \mathbb{E} [|Y_\infty| \mathbf{1}_{\{Y_\infty < -1/\delta\}}] \right\}^2 - \mu\delta^{-1}\mathbb{E} [|Y_\infty| \mathbf{1}_{\{Y_\infty < -1/\delta\}}] - \left[c_A + \frac{3}{2}(1 - \mu)\delta^2 \right] \leq 0.$$

Solving the above quadratic inequality gives

$$\begin{aligned} \mathbb{E} [|Y_\infty| \mathbf{1}_{\{Y_\infty < -1/\delta\}}] & \leq \frac{1}{4}\mu\delta^{-1} + \frac{1}{4}\sqrt{\mu^2\delta^{-2} + 8 \left[c_A + \frac{3}{2}(1 - \mu)\delta^2 \right]} \\ & \leq \frac{1}{2}\mu\delta^{-1} + \frac{1}{2}\sqrt{2c_A + 3(1 - \mu)\delta^2} \\ & \leq \frac{1}{2}\mu\delta^{-1} + c_A + \frac{3}{2}(1 - \mu)\delta^2, \end{aligned} \quad (4.45)$$

110 where for the second inequality we use the fact that for any two non-negative real numbers y and z , $\sqrt{y+z} \leq \sqrt{y} + \sqrt{z}$, and for the last inequality we use $c_A \geq 1$.

Substituting (4.45) back into (4.44), we have

$$\begin{aligned} & 2\mathbb{E} [Y_\infty^2 \mathbf{1}_{\{Y_\infty < -1/\delta\}}] + \mathbb{E} [Y_\infty^2 \mathbf{1}_{\{Y_\infty \in [-1/\delta, -\zeta]\}}] + |\zeta| \mathbb{E} [Y_\infty \mathbf{1}_{\{Y_\infty \geq -\zeta\}}] \\ & \leq c_A + \frac{3}{2}(1 - \mu)\delta^2 + \mu\delta^{-1} \left[c_A + \frac{3}{2}(1 - \mu)\delta^2 \right] + \frac{1}{2}\mu^2\delta^{-2}. \end{aligned} \quad (4.46)$$

Then, applying Jensen's inequality to (4.46) implies that

$$\begin{aligned} \mathbb{E} [|Y_\infty| \mathbf{1}_{\{Y_\infty \in [-1/\delta, -\zeta]\}}] & \leq \sqrt{\mathbb{E} [Y_\infty^2 \mathbf{1}_{\{Y_\infty \in [-1/\delta, -\zeta]\}}]} \\ & \leq \sqrt{1 + \mu\delta^{-1}} \sqrt{c_A + \frac{3}{2}(1 - \mu)\delta^2} + \sqrt{\frac{1}{2}\mu^2\delta^{-2}} \\ & \leq \left(1 + \mu^{1/2}\delta^{-1/2} \right) \left[c_A + \frac{3}{2}(1 - \mu)\delta^2 \right] + \frac{\sqrt{2}}{2}\mu\delta^{-1}. \end{aligned} \quad (4.47)$$

Finally,

$$\mathbb{E} [Y_\infty \mathbf{1}_{\{Y_\infty \geq -\zeta\}}] \leq \frac{1}{|\zeta|} \left\{ (1 + \mu\delta^{-1}) \left[c_A + \frac{3}{2}(1 - \mu)\delta^2 \right] + \frac{1}{2}\mu^2\delta^{-2} \right\}. \quad (4.48)$$

Adding up (4.45), (4.47), and (4.48) proves Lemma 4.2. \square

Proof of Lemma 4.3. We start with (4.12).

From the form of $a(x)$ in (3.2), (4.12) is obviously true for $x \leq -1/\delta$ and $x \geq -\zeta$.

When $x \in [-1/\delta, -\zeta]$,

$$a(x) = \mu (c_A - \mu + \delta(1 - \mu)x + \mu x^2) = \mu \left\{ \mu (x - x_0)^2 + (c_A - \mu) - \mu x_0^2 \right\},$$

115 where $x_0 = -\frac{\delta(1-\mu)}{2\mu}$.

If $x_0 \leq -1/\delta$, $a(x)$ is increasing with x over the interval $[-1/\delta, -\zeta]$. Thus

$$\min_{x \in [-1/\delta, -\zeta]} a(x) \geq a(-1/\delta) > (c_A - 1)\mu.$$

Otherwise, we have $\delta^2 < \frac{2\mu}{1-\mu}$.

$$\begin{aligned} \min_{x \in [-1/\delta, -\zeta]} a(x) &= a(x_0) = \mu(c_A - \mu) - \mu^2 \frac{\delta^2(1 - \mu)^2}{4\mu^2} \\ &> \mu(c_A - \mu) - \frac{1}{4}(1 - \mu)^2 \frac{2\mu}{1 - \mu} \\ &= \mu(c_A - 1/2 - \mu/2) \\ &> (c_A - 1)\mu, \end{aligned}$$

where to get the last inequality we use $\mu < 1$.

In this way we have established (4.12).

Next we proceed to (4.13). Note that

$$a'(x) = \mu [2\mu x + \delta(1 - \mu)] \mathbf{1}_{\{x \in (-1/\delta, -\zeta]\}}. \quad (4.49)$$

Hence

$$\frac{|xa'(x)|}{a(x)} = \frac{|2\mu x^2 + \delta(1 - \mu)x|}{\mu x^2 + \delta(1 - \mu)x + c_A - \mu} \mathbf{1}_{\{x \in (-1/\delta, -\zeta]\}}. \quad (4.50)$$

Consider the function

$$g_1(x) = 2 [\mu x^2 + \delta(1 - \mu)x + c_A - \mu] - [2\mu x^2 + \delta(1 - \mu)x] = \delta(1 - \mu)x + 2(c_A - \mu).$$

When $x \in [-1/\delta, -\zeta]$,

$$g_1(x) \geq -(1 - \mu) + 2(c_A - \mu) = (c_A - \mu) + (c_A - 1) > 0, \quad (4.51)$$

where to get the last inequality we use $\mu < 1$ and $c_A > 1$.

Consider another function

$$\begin{aligned} g_2(x) &= a [\mu x^2 + \delta(1 - \mu)x + c_A - \mu] + [2\mu x^2 + \delta(1 - \mu)x] \\ &= (a + 2)\mu x^2 + (a + 1)\delta(1 - \mu)x + a(c_A - \mu), \end{aligned}$$

120 where a is a positive constant to be determined later.

When $x \in [-1/\delta, -\zeta]$,

$$g_2(x) \geq -(a + 1)(1 - \mu) + a(c_A - \mu) = (c_A - 1)a - (1 - \mu).$$

Taking $a = \frac{1-\mu}{c_A-1}$, we have

$$g_2(x) \geq 0, \quad x \in [-1/\delta, -\zeta]. \quad (4.52)$$

Applying (4.51) and (4.52) to (4.50) proves (4.13).

Now we deal with (4.14).

From (3.2) and (4.49),

$$\frac{|a'(x)|}{a(x)} = \frac{|2\mu x + \delta(1 - \mu)|}{\mu x^2 + \delta(1 - \mu)x + (c_A - \mu)} \mathbf{1}_{\{x \in (-1/\delta, -\zeta]\}}. \quad (4.53)$$

Consider the function

$$\begin{aligned} g_3(x) &= a [\mu x^2 + \delta(1 - \mu)x + (c_A - \mu)] - [2\mu x + \delta(1 - \mu)] \\ &= a\mu x^2 + [\delta a(1 - \mu) - 2\mu]x + a(c_A - \mu) - \delta(1 - \mu), \end{aligned}$$

where a is a positive constant to be determined later.

For $\mu \in (0, 1)$ such that $\delta a(1 - \mu) - 2\mu \geq 0$, when $x \in [-1/\delta, -\zeta]$,

$$\begin{aligned} g_3(x) &\geq -a(1 - \mu) + 2\mu\delta^{-1} + a(c_A - \mu) - \delta(1 - \mu) \\ &\geq -a(1 - \mu) + a(c_A - \mu) - (1 - \mu) \\ &= (c_A - 1)a - (1 - \mu), \end{aligned}$$

where for the second inequality we used $\delta \leq 1$.

For $\mu \in (0, 1)$ such that $\delta a(1 - \mu) - 2\mu < 0$, let $x_0 = \frac{2\mu - \delta a(1 - \mu)}{2a\mu} > 0$.

$$\begin{aligned} g_3(x) &= a\mu(x - x_0)^2 + a(c_A - \mu) - \delta(1 - \mu) - a\mu x_0^2 \\ &\geq a(c_A - \mu) - \delta(1 - \mu) - a\mu \left(\frac{2\mu}{2a\mu} \right)^2 \\ &\geq a(c_A - \mu) - (1 - \mu) - \mu/a \\ &\geq [(c_A - 1)a - 1] + \mu(1 - 1/a). \end{aligned}$$

Taking $a = \frac{1}{c_A - 1} \vee 1$, we have that

$$g_3(x) \geq 0, \quad x \in [-1/\delta, -\zeta]. \quad (4.54)$$

Consider another function

$$\begin{aligned} g_4(x) &= a[\mu x^2 + \delta(1 - \mu)x + (c_A - \mu)] + [2\mu x + \delta(1 - \mu)] \\ &= a\mu x^2 + [\delta a(1 - \mu) + 2\mu]x + a(c_A - \mu) + \delta(1 - \mu), \end{aligned}$$

¹²⁵ where a is a positive constant to be determined later.

Denote $x_0 = -\frac{\delta a(1 - \mu) + 2\mu}{2a\mu}$. If $x_0 \leq -1/\delta$, $g_4(x)$ is increasing with x over the interval $[-1/\delta, -\zeta]$.

Then, for $x \in [-1/\delta, -\zeta]$,

$$\begin{aligned} g_4(x) &\geq g_4(-1/\delta) \\ &= a\mu\delta^{-2} - a(1 - \mu) - 2\mu\delta^{-1} + a(c_A - \mu) + \delta(1 - \mu) \\ &\geq (a - 2)\mu\delta^{-1} + (c_A - 1)a. \end{aligned}$$

Otherwise, $x_0 > -1/\delta$, that is,

$$\delta a(1 - \mu) + 2\mu < 2a\mu\delta^{-1}. \quad (4.55)$$

Then,

$$\begin{aligned}
g_4(x) &= a\mu(x - x_0)^2 + a(c_A - \mu) + \delta(1 - \mu) - a\mu x_0^2 \\
&\geq a(c_A - \mu) + \delta(1 - \mu) - a\mu \left(\frac{\delta a(1 - \mu) + 2\mu}{2a\mu} \right)^2 \\
&> a(c_A - \mu) + \delta(1 - \mu) - \frac{1}{2}\delta^{-1} [\delta a(1 - \mu) + 2\mu] \\
&\geq a(c_A - \mu) - \frac{1}{2}\delta^{-1} [\delta a(1 - \mu) + 2\mu] \\
&= c_A a - \frac{1}{2}(1 + \mu)a - \mu\delta^{-1} \\
&\geq (c_A - 1)a - \mu\delta^{-1},
\end{aligned}$$

where to get the third line we use (4.55).

Taking $a = \frac{\mu\delta^{-1}}{c_A - 1} \vee 2$, we have that

$$g_4(x) \geq 0, \quad x \in [-1/\delta, -\zeta]. \quad (4.56)$$

Applying (4.54) and (4.56) to (4.53), we get that

$$\frac{|a'(x)|}{a(x)} \leq \left(\frac{\mu\delta^{-1}}{c_A - 1} \vee \frac{1}{c_A - 1} \vee 2 \right) \mathbb{1}_{\{x \in (-1/\delta, -\zeta]\}}. \quad (4.57)$$

Using (4.57) along with (4.11) in Lemma 4.2 proves (4.14).

Finally we approach (4.15).

Note that when $x \leq -1/\delta$,

$$|r'(x)a(x)| = 2\mu, \quad (4.58)$$

and when $x > -\zeta$,

$$|r'(x)a(x)| = 0. \quad (4.59)$$

When $x \in (-1/\delta, -\zeta]$,

$$\begin{aligned}
r'(x)a(x) &= \frac{2}{a(x)} [b'(x)a(x) - b(x)a'(x)] \\
&= -2\mu + 2\mu \frac{xa'(x)}{a(x)}.
\end{aligned}$$

Hence

$$|r'(x)a(x)| \leq 2\mu \left(1 + \frac{|xa'(x)|}{a(x)} \right).$$

Applying (4.13) to the right side of the inequality above, along with (4.58) and (4.59), proves (4.15). \square

References

- [1] J. Feng, P. Shi, Steady-state diffusion approximations for discrete-time queue in hospital inpatient flow management, working paper (2016).
- [2] J. Feng, P. Shi, Online supplement to steady-state diffusion approximations for discrete-time queue in hospital inpatient flow management, online supplement (2016).
- [3] E. W. Weisstein, [Poisson distribution](http://mathworld.wolfram.com/PoissonDistribution.html), from MathWorld—A Wolfram Web Resource.
URL <http://mathworld.wolfram.com/PoissonDistribution.html>
- [4] E. W. Weisstein, [Binomial distribution](http://mathworld.wolfram.com/BinomialDistribution.html), from MathWorld—A Wolfram Web Resource.
URL <http://mathworld.wolfram.com/BinomialDistribution.html>
- [5] J. G. Dai, P. Shi, [A two-time-scale approach to time-varying queues in hospital inpatient flow management](http://papers.ssrn.com/sol3/papers.cfm?abstract_id=2489533), Operations Research, *forthcoming*.
URL http://papers.ssrn.com/sol3/papers.cfm?abstract_id=2489533
- [6] A. Braverman, J. G. Dai, [High order steady-state diffusion approximation of the Erlang-C system](http://arxiv.org/abs/1602.02866), submitted for publication (2016).
URL <http://arxiv.org/abs/1602.02866>
- [7] S. P. Meyn, R. L. Tweedie, Stability of Markovian processes III: Foster-Lyapunov criteria for continuous time processes, Adv. Appl. Probab. 25 (1993) 518–548.